NORMAND PETTERSEN

Évaluation du potentiel humain dans les organisations

Élaboration et validation d'instruments de mesure

Presses de l'Université du Québec

Évaluation du potentiel humain dans les organisations

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

PRESSES DE L'UNIVERSITÉ DU QUÉBEC
Le Delta I, 2875, boul. Laurier, bureau 450
Sainte-Foy (Québec) G1V 2M2
Téléphone : (418) 657-4399 · Télécopieur : (418) 657-2096
Courriel : puq@puq.uquebec.ca · Internet : www.puq.uquebec.ca

Distribution :

CANADA et autres pays

DISTRIBUTION DE LIVRES UNIVERS S.E.N.C.
845, rue Marie-Victorin, Saint-Nicolas (Québec) G7A 3S8
Téléphone : (418) 831-7474 / 1-800-859-7474 • Télécopieur : (418) 831-4021

FRANCE

DIFFUSION DE L'ÉDITION QUÉBÉCOISE
30, rue Gay-Lussac, 75005 Paris, France
Téléphone : 33 1 43 54 49 02
Télécopieur : 33 1 43 54 39 15

SUISSE

SERVIDIS SA
5, rue des Chaudronniers, CH-1211 Genève 3, Suisse
Téléphone : 021 803 26 26
Télécopieur : 021 803 26 29

DANGER
LE PHOTOCOPILLAGE TUE LE LIVRE

La *Loi sur le droit d'auteur* interdit la reproduction des œuvres sans autorisation des titulaires de droits. Or, la photocopie non autorisée – le « photocopillage » – s'est généralisée, provoquant une baisse des ventes de livres et compromettant la rédaction et la production de nouveaux ouvrages par des professionnels. L'objet du logo apparaissant ci-contre est d'alerter le lecteur sur la menace que représente pour l'avenir de l'écrit le développement massif du « photocopillage ».

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

NORMAND PETTERSEN

Évaluation du potentiel humain dans les organisations

Élaboration et validation d'instruments de mesure

2002

Presses de l'Université du Québec
Le Delta I, 2875, boul. Laurier, bur. 450
Sainte-Foy (Québec) Canada G1V 2M2

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Données de catalogage avant publication (Canada)

Pettersen, Normand

Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure

Comprend des réf. bibliogr.

ISBN 2-7605-1051-4

1. Personnel – Évaluation. 2. Ressources humaines. 3. Potentiel humain (Psychologie). I. Titre.

HF5549.5.R3P48 2000 658.3'125 C00-941126-7

Nous reconnaissons l'aide financière du gouvernement du Canada par l'entremise du Programme d'aide au développement de l'industrie de l'édition (PADIÉ) pour nos activités d'édition.

Révision linguistique : GISLAINE BARRETTE

Mise en pages : CARACTÉRA PRODUCTION GRAPHIQUE INC.

Couverture : RICHARD HODGSON

1 2 3 4 5 6 7 8 9 PUQ 2002 9 8 7 6 5 4 3 **2** 1

Tous droits de reproduction, de traduction et d'adaptation réservés
© 2000 Presses de l'Université du Québec

Dépôt légal – 3e trimestre 2000
Bibliothèque nationale du Québec / Bibliothèque nationale du Canada
Imprimé au Canada

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

TABLE DES MATIÈRES

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

LISTE DES TABLEAUX

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

LISTE DES FIGURES

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca

Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

REMERCIEMENTS

Ce livre est le résultat de la collaboration et de la générosité de plusieurs personnes. Je remercie d'abord mes étudiants dont les commentaires et les questions (ou simplement leur air étonné…) ont été de puissants catalyseurs dans le processus de clarification des idées. Certains magistrats, à leur insu, ont produit le même effet.

Des représentants du monde universitaire et professionnel ont contribué à enrichir ce livre. Je remercie Line Cardinal, de l'Université du Québec à Montréal, pour ses commentaires judicieux sur une version préliminaire du manuscrit. Jacques Barrette, de l'Université d'Ottawa, a révisé la presque totalité du manuscrit ; son soutien réconfortant s'est fait sentir tout au long de cette aventure. Il y a aussi Jean Bouchard, du Centre hospitalier universitaire de Québec, Jean Herickx, professeur en retraite active, Jean Lortie, de l'Institut de police du Québec, Marc-André Verrette, du Groupe ressources (DGO) et François Bernatchez, consultant, qui ont offert spontanément de revoir d'importantes parties du manuscrit. Je suis également reconnaissant à mes collègues Camille Carrier et Bruno Fabi pour leurs suggestions et leurs encouragements.

Je souligne le professionnalisme et la souplesse du personnel des Presses de l'Université du Québec, sans oublier la compétence de Suzanne Hamel, de l'Université du Québec à Trois-Rivières.

J'aimerais exprimer ma gratitude à Louise, ma conjointe, pour avoir accepté de bonne grâce les heures dérobées à la famille et pour avoir dit que je travaillais fort à ceux qui me croyaient en congé.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

INTRODUCTION

La qualité du personnel est un facteur clé dans le succès d'une organisation ou la réussite d'un projet ; l'avantage concurrentiel d'une entreprise passe par l'immense potentiel offert par les femmes et les hommes qui la composent. Et c'est loin d'être terminé, compte tenu de la concurrence devenue planétaire, de la complexité croissante des technologies et de la gestion toujours plus décentralisée. Pour bénéficier de ce potentiel humain, les décisions d'ordre stratégique et opérationnel doivent tenir compte de ce que chaque personne a de mieux à offrir, que ce soit lors de la sélection et de l'affectation du personnel, de la gestion et de l'évaluation du rendement, de la planification et du développement de la relève, de l'élaboration et du suivi de la formation ou encore lors de la conception et de l'application des diverses formes de rémunération au mérite.

Ce constat n'est pas très original ; de nombreux témoins de l'activité économique le répètent continuellement. Néanmoins, tenir compte des différences individuelles n'est pas simple en pratique, à commencer par les défis constants que posent la mesure et l'évaluation des personnes, de leur rendement et de leurs caractéristiques. Par exemple, comment s'y prendre pour élaborer un examen de compétences, une entrevue de sélection, un exercice de mise en situation, une grille d'évaluation du rendement ou tout autre instrument de mesure ? Quels sont les meilleurs outils en sélection du personnel, ceux qui prédisent le mieux le rendement des candidats une fois en poste ? Est-ce que les tests psychométriques mesurant l'intelligence générale sont supérieurs à l'entrevue ? Comment établir le profil du candidat idéal qui correspond aux exigences de l'emploi ? Comment savoir si un outil de sélection (p. ex., les résultats à une entrevue) est

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

vraiment un gage de réussite dans l'emploi ? Quelle est la marge d'erreur et comment la réduire ? Lors d'une procédure de dotation interne, est-ce rentable pour l'organisation de recourir à plusieurs instruments de mesure requérant trois journées entières d'évaluation ? Lorsqu'il y a litige, comment peut-on démontrer la pertinence des outils de sélection par rapport aux exigences de l'emploi ? Dans un processus d'évaluation du rendement, comment obtenir des résultats équitables ? Comment peut-on en améliorer la fiabilité et ainsi accroître la confiance des participants ? Mais d'abord, qu'est-ce qu'une bonne mesure du rendement ? Que doit-elle inclure ?

Voilà un échantillon des problèmes qui préoccupent les gestionnaires et les professionnels en ressources humaines à un moment ou à un autre de leur travail. Bien entendu, les résoudre relève avant tout de la gestion et exige de bien circonscrire les paramètres économiques, technologiques, légaux, politiques et sociaux de l'organisation. Cependant, les solutions ne pourront être pleinement efficaces si elles ne prennent appui sur les connaissances et les pratiques de la psychométrie. Ce domaine, appelé aussi la mesure et l'évaluation, est vaste et complexe ; avec ses concepts abstraits et ses fondements statistiques, il en rebute plusieurs. Pourtant, il permet d'adopter une approche plus rigoureuse de la gestion des ressources humaines lors de l'évaluation des individus, de leurs compétences ou de leur rendement.

Le but du présent ouvrage est de montrer, dans un langage accessible, comment élaborer et valider de façon scientifique les nombreux instruments de mesure utilisés en gestion des ressources humaines. On pense bien sûr aux instruments de sélection ou de qualification comme l'entrevue structurée, les simulations, les tests et les examens, mais il y a aussi les instruments servant à évaluer le rendement des personnes au travail ou à mesurer les apprentissages après formation. Ces instruments doivent être simples, rigoureux et défendables en cas de litige. Il faut également savoir comment en estimer la valeur pour l'organisation et pouvoir en justifier l'usage.

On retrouve aussi dans cet ouvrage des informations souvent indispensables en gestion des ressources humaines, en relations industrielles ou en psychologie industrielle/organisationnelle, notamment : 1) des critères servant à choisir et à déterminer la valeur tangible d'un instrument de mesure, 2) une analyse comparative des méthodes de sélection du personnel, 3) des moyens pour contrôler

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

les erreurs de mesure et augmenter la fiabilité (fidélité) des résultats, 4) un processus d'analyse et de description de tâches adapté à l'élaboration d'instruments de mesure, 5) des typologies portant sur les caractéristiques humaines et servant à identifier les qualités requises pour un emploi, 6) les qualités d'une bonne mesure du rendement, 7) une présentation simple des fondements statistiques et de leur utilisation concrète, 8) des références bibliographiques abondantes, à jour et dignes de confiance.

Orienté vers la pratique et comportant de multiples exemples du monde réel, *Évaluation du potentiel humain dans les organisations* s'adresse au professionnel, gestionnaire ou consultant, pour qui la rigueur est un gage d'efficacité. Appuyé sur une analyse approfondie de la documentation scientifique récente, l'ouvrage se veut aussi un outil de référence pour le spécialiste de la mesure et de l'évaluation préoccupé par les applications au monde du travail. En outre, il trouvera dans les notes en bas de page des explications additionnelles portant sur des aspects plus complexes ou plus avancés. Finalement, cet ouvrage peut se révéler un compagnon des plus utiles en cas de litige ou de poursuite devant les tribunaux.

Plusieurs praticiens pourront penser que ce livre ne s'adresse pas à eux ; trop compliqués, les instruments de mesure sont une affaire de spécialistes, se diront-ils. Avant de clore le débat cependant, ils devraient considérer les faits suivants. Premièrement, ils n'auront pas toujours le temps de s'en remettre à un spécialiste chaque fois qu'un événement les pressera d'agir. Deuxièmement, les spécialistes en mesure qui sont également expérimentés en gestion des ressources humaines se font rares : il y en a très peu de formés dans les universités et la plupart des organisations ne disposent pas de l'expertise pour assurer elles-mêmes la relève. Troisièmement, ce livre a précisément pour objectif de rendre accessible ce domaine et de l'appliquer au contexte de la gestion des ressources humaines. Finalement, il faut peut-être se demander si la rigueur est facultative en gestion des ressources humaines ou si elle est un devoir professionnel.

L'ouvrage comprend sept chapitres. Le premier montre comment déterminer la valeur des instruments de mesure et leur utilité économique. Les approches proposées vont de la simple comparaison des avantages et des inconvénients aux calculs complexes du retour

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

sur l'investissement. On y apprend entre autres que l'usage des instruments de mesure, du moins en sélection du personnel, est généralement très rentable pour l'organisation.

Le deuxième chapitre porte sur la validation basée sur le contenu de l'instrument de mesure. Particulièrement utile en gestion des ressources humaines, cette forme de validation vise à démontrer la pertinence du contenu d'un instrument pour un usage donné. Après une introduction au concept de validité pris au sens large, ce chapitre présente en détail la validation basée sur le contenu et ses nombreuses implications pour le monde du travail.

La validation basée sur la relation avec d'autres variables vérifie, auprès d'un échantillon de candidats, jusqu'à quel point les résultats obtenus à un instrument de mesure sont reliés à des variables extérieures à cet instrument, notamment au rendement au travail. Le troisième chapitre explique comment conduire une telle étude de validation et comment juger si la valeur du coefficient de validité obtenue est suffisante. La mesure du rendement au travail et les retombées des méta-analyses pour la gestion des ressources humaines font également partie de l'exposé.

La fidélité et le contrôle des erreurs de mesure sont l'objet du quatrième chapitre. On y traite des diverses sources d'erreurs dans un instrument de mesure qui affectent la précision et la stabilité des résultats. Les moyens à prendre pour contrôler et diminuer ces erreurs, si dommageables en gestion des ressources humaines, sont passés en revue. La suite du chapitre aborde le concept de fidélité et en énumère les usages pratiques.

Les cinquième et sixième chapitres sont consacrés à l'élaboration d'instruments de mesure utilisés en gestion des ressources humaines. L'approche d'élaboration proposée s'appuie principalement sur la stratégie de validation basée sur le contenu ; centrée sur le contenu de l'emploi à effectuer, elle s'oriente naturellement vers des instruments de type 1) entrevues structurées, 2) mises en situation et analyse de cas (*In-Basket*, exercice de prise de décision en groupe, etc.), 3) examens de connaissances, 4) échantillons de travail (test de dactylographie, examen pratique de soudure, de conduite d'un véhicule, etc.), ou 5) grille d'évaluation du rendement.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Le septième chapitre porte sur les fondements statistiques utilisés dans la mesure et l'évaluation des individus et de leurs caractéristiques. C'est un survol de notions élémentaires, dont plusieurs sont essentielles à la compréhension des concepts faisant l'objet des chapitres précédents. Le lecteur qui n'est pas familier avec ces notions serait avisé de lire ce chapitre en premier ou, à tout le moins, de s'y référer au besoin.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

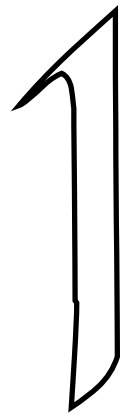
Tous droits de reproduction, de traduction ou d'adaptation réservés

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

C H A P I T R E



VALEUR DES INSTRUMENTS DE MESURE ET LEUR UTILITÉ ÉCONOMIQUE

Comment établir la valeur d'un instrument de mesure – test psychométrique, examen de connaissances, entrevue, simulation, etc. – utilisé en gestion des ressources humaines ? Quelle perspective doit adopter l'utilisateur qui veut choisir, élaborer ou justifier l'usage d'un tel instrument dans une situation donnée ? Est-il judicieux pour une organisation de recourir à ces instruments pour gérer ses ressources humaines ? Pour répondre à ces questions et rendre compte de la valeur pratique d'un instrument de mesure, il faut se placer dans un contexte de gestion, où les enjeux sont innombrables, à géométrie variable selon les acteurs concernés et surtout propres à chacune des situations. Il ne s'agit pas seulement de juger la valeur scientifique de l'instrument de mesure, mais aussi d'adopter une perspective de prise de décision et de vérifier si l'emploi de l'instrument de mesure est justifié par les multiples enjeux envisagés.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Plan du chapitre. Ce chapitre aborde la façon de déterminer la valeur globale d'un instrument de sélection lorsqu'il est appliqué dans une situation donnée. Les approches proposées vont de la simple comparaison des avantages et des inconvénients aux calculs complexes du retour sur l'investissement. Le chapitre est divisé en deux sections : la première présente un ensemble de critères pouvant servir à dresser le portrait des avantages et des inconvénients découlant de l'usage d'un instrument de sélection ; la seconde, plus quantitative, porte sur des méthodologies qui cherchent à calculer l'apport d'un outil de sélection tantôt au regard de l'augmentation du rendement, tantôt au regard du gain économique. Tout au long de ce chapitre, le contexte d'analyse est celui de la sélection du personnel. Toutefois, le lecteur n'aura pas de difficulté à transposer la démarche, du moins dans son essence, aux autres pratiques de gestion des ressources humaines.

Ce qu'il faut en retenir. Ce premier chapitre sert de cadre général pour les principaux concepts présentés dans cet ouvrage : le concept de validité en général, la validité basée sur le contenu de l'instrument, la validité basée sur la relation avec d'autres variables, la fidélité, etc., sans compter quelques notions statistiques. Pour l'instant, il n'est pas essentiel d'avoir une connaissance complète de ces matières ; le lecteur n'a donc pas à s'inquiéter outre mesure si certains aspects lui échappent. Il lui faut plutôt se concentrer sur l'idée générale de ce chapitre qui se résume aux deux messages suivants.

Le premier est que la valeur d'un instrument de mesure n'est pas absolue, mais dépend des paramètres de la situation dans laquelle il est utilisé. Dès lors, l'introduction ou le maintien d'un instrument de mesure doivent faire l'objet d'une décision éclairée par un bilan des avantages et des inconvénients. Le ratio « retour sur investissement » ne doit pas non plus être écarté comme soutien à la décision ; il est possible de nos jours d'en faire rapidement une estimation. Quant au deuxième message, il concerne la rentabilité des instruments de mesure en gestion des ressources humaines. Les résultats de nombreuses recherches dépassent les espoirs les plus grands, du moins pour les outils de sélection du personnel : leur usage est très rentable. En outre, ce sont les outils les plus rigoureux qui semblent les plus avantageux : la rigueur n'est plus un luxe, mais un investissement !

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

CRITÈRES POUR JUGER LA VALEUR D'UN INSTRUMENT DE SÉLECTION

La validité est la qualité première d'un instrument de mesure ; elle indique jusqu'à quel point l'instrument parvient à mesurer ce qu'il est censé mesurer ou à prédire ce qu'il est censé prédire. Par exemple, un examen d'anglais est valide s'il mesure vraiment la compétence en anglais, et seulement cette compétence. Une entrevue de sélection est valide si elle permet de déterminer lesquels, parmi les candidats, auront le meilleur rendement une fois en poste. Si la validité est nulle ou pratiquement inexistante, un instrument n'a pas de valeur et son utilisation peut difficilement se justifier. En revanche, si l'instrument a démontré un certain degré de validité, il peut être utile à l'organisation. Mais une question difficile attend alors l'usager : à partir de quel niveau de validité peut-il considérer qu'un instrument de sélection est valable ?

VALIDITÉ : QUALITÉ ESSENTIELLE EN THÉORIE, MAIS INSUFFISANTE EN PRATIQUE

D'un point de vue **théorique**, il est possible d'apprécier le niveau de validité. En sélection du personnel, par exemple, la validité renvoie souvent à la relation entre les résultats des candidats obtenus à l'instrument de sélection et leur performance en emploi une fois embauchés[1]. L'instrument de mesure est considéré valide si les candidats qui obtiennent les meilleurs résultats à l'instrument sont aussi les employés les plus performants, et vice versa. Cette relation entre l'instrument de mesure et le rendement est généralement quantifiée par un indice statistique, le coefficient de corrélation (dans les circonstances, appelé coefficient de validité critériée), dont la valeur absolue varie de 0,00 à 1,00. Plus la validité est élevée, plus le coefficient augmente (voir chapitre 7). Par exemple, les meilleurs

1. Cette façon de concevoir la validité découle d'une approche appelée « validation basée sur la relation avec d'autres variables », parce que les résultats obtenus à l'instrument de mesure sont mis en relation avec une autre variable, en l'occurrence le rendement au travail ; cette autre variable sert alors de critère pour vérifier la validité de l'instrument. Nous verrons plus loin qu'il existe d'autres façons d'estimer la validité.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

instruments de mesure utilisés en sélection ont des validités avoisinant parfois des coefficients de 0,50 à 0,60. Est-ce une validité suffisante pour l'organisation ?

Interpréter rigoureusement la valeur du coefficient de validité est beaucoup plus exigeant qu'il n'y paraît. Cette question, pourtant simple, possède des ramifications méthodologiques nombreuses et fort complexes ; un chapitre presque complet traite en détail chacun de ces aspects (voir chapitre 3). Pour l'instant, disons qu'il est usuel de simplifier l'interprétation du coefficient de validité en calculant le coefficient de détermination, qui est égal au coefficient de validité élevé au carré ; le coefficient devient alors un pourcentage. Par exemple, un indice de validité de 0,50 pour une entrevue de sélection indique que les résultats des candidats à cette entrevue sont reliés à 25 % à leur rendement au travail une fois en poste (soit $0{,}50^2 = 0{,}25$ ou 25 %). Autrement dit, l'entrevue permet de prédire 25 % du rendement de ces candidats.

Inconvénients pratiques. Pour l'homme de la rue, pour le gestionnaire et même pour un spécialiste en psychométrie, l'indice de validité ne permet pas d'apprécier la valeur globale d'un instrument. Aussi important que puisse être la validité, la valeur d'un instrument de mesure ne peut se résumer à cette seule qualité sans comporter plusieurs inconvénients. Le premier inconvénient est que le coefficient de validité est un **concept abstrait**, pour ne pas dire insaisissable, pour la plupart des gens. Tenter de convaincre à coup de coefficients de corrélation pourra au mieux susciter un respect poli à l'égard de l'argumentation scientifique, au pire, beaucoup de scepticisme.

Le calcul du coefficient de détermination (corrélation élevée au carré), qui transforme la validité en pourcentage de prédiction, a permis de rendre l'indice de validité un peu plus concret, non sans entraîner un deuxième inconvénient de taille. C'est qu'une fois transformée en coefficient de détermination, la validité semble **plus faible**, voire ridicule. Par exemple, un coefficient de validité de 0,50 peut sembler intéressant aux yeux d'une personne non férue de statistiques, mais une fois traduite en une capacité de prédiction de 25 % (soit 0,50 élevé au carré), cela risque de décevoir. Pourtant, nous serions en présence d'un très bon instrument de mesure, en comparaison de ce qui existe déjà. On imagine aussitôt l'effet produit par un

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

coefficient de 0,30 qui serait ramené à un maigre 9 % de prédiction. Certains penseront que c'est le néant ! Ces mathématiques ont vite créé l'impression que la validité doit être substantiellement élevée pour qu'un instrument de sélection soit vraiment utile pour l'organisation.

Mais voilà, nous ne sommes pas en mathématiques, avec ses paramètres absolus : nous sommes **en pratique**, où tout est relatif. Or, la valeur globale d'un instrument varie en fonction de divers facteurs de la situation, que les coefficients de corrélation ou de détermination sont **impuissants** à reconnaître (Schneider et Schmitt, 1986)[2]. Par exemple, un instrument de sélection est moins utile lorsque le marché du travail regorge de candidats très qualifiés, simplement parce qu'il y a très peu de candidatures non qualifiées à éliminer ; que l'instrument soit utilisé ou non, la qualité du personnel embauché sera élevée. Ce troisième inconvénient constitue certainement la plus grande limite du coefficient de validité.

Solution. Alors, à partir de quel niveau un instrument de sélection est-il suffisamment valide pour être utilisé dans une situation en particulier ? Du point de vue organisationnel, il faut se placer dans un contexte où un instrument de sélection n'a pas à être parfait, mais doit contribuer à améliorer les décisions et présenter plus d'avantages que d'inconvénients. C'est pourquoi *un instrument est suffisamment valide à partir du moment où son application engendre des améliorations qui dépassent les coûts et les autres répercussions négatives*. Par exemple, un outil dont l'utilisation entraîne des coûts plus élevés que l'augmentation de rendement obtenue ne serait pas très valable pour l'organisation. En résumé, la validité est une qualité essentielle, mais non suffisante. La valeur globale d'un outil ou d'une méthode de sélection dépend également des nombreuses retombées qui découlent de l'application de cet outil ou de cette méthode (Guion, 1998).

2. Ne tenant pas compte de la spécificité de la situation, la corrélation traite les erreurs de prédiction de la même manière, qu'elles soient positives ou négatives. Par conséquent, un candidat qui a un rendement supérieur à celui prévu par l'instrument de sélection constitue une erreur de prédiction au même titre qu'un autre qui affiche un rendement inférieur à la prévision.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

AUTRES CRITÈRES D'ÉVALUATION À CONSIDÉRER

Les retombées que peut avoir un outil de sélection sont multiples : augmentation du rendement des employés, de la qualité du travail effectué, de la fiabilité, de la sécurité au travail ou baisse des coûts de main-d'œuvre, du temps de formation, du taux de roulement, de l'absentéisme et autres retombées reliées directement au processus de sélection. Ces conséquences ne peuvent advenir que si les instruments de sélection possèdent un certain niveau de validité. D'autres conséquences, moins dépendantes de la validité, sont aussi à prendre en considération comme la possibilité de contestations légales et de plaintes de discrimination, la capacité de se défendre en cas de litige, la mauvaise publicité, la perception d'iniquité chez les employés, etc. En réalité, on peut s'intéresser à une multitude de conséquences résultant de l'application d'une méthode de sélection, que ce soit du point de vue de l'organisation, de ses employés ou même de la société en général. Par conséquent, *la valeur d'un instrument de mesure dépend de sa capacité à maximiser les avantages souhaités et à minimiser les inconvénients redoutés*. Dans ce contexte, l'instrument absolu n'existe pas ; il faut plutôt voir lequel est le plus avantageux pour l'usager en fonction des objectifs poursuivis, tout en tenant compte des caractéristiques de la situation.

Procéder ainsi à l'appréciation d'un instrument du point de vue de ses avantages et de ses inconvénients exige que l'on choisisse des critères d'évaluation, choix déterminant quant à l'issue de la démarche (Boudreau, 1991). Le nombre de critères possibles est pratiquement illimité, à l'instar de la multitude des conséquences, des avantages ou des inconvénients qui peuvent découler de l'application d'une méthode de sélection. Cependant, en plus de la validité, il y a des critères qui sont plus ou moins des incontournables pour l'organisation. Quels sont ces critères d'évaluation ?

Coûts directs d'utilisation. D'abord, il y a les coûts engendrés directement par l'application d'un outil de sélection, incluant s'il y a lieu son développement et sa mise à jour. À titre d'illustration, voyons une analyse comparative de diverses méthodes de sélection qui intègre simultanément validité et coûts d'utilisation (tableau 1.1). Les valeurs attribuées sont approximatives et ne peuvent être considérées comme des indices rigoureusement sûrs par rapport à la réalité, bien que celles sur la validité s'appuient sur de très nombreuses

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 1.1
ANALYSE COMPARATIVE DE DIVERSES MÉTHODES DE SÉLECTION DU PERSONNEL EN FONCTION DE LA VALIDITÉ CRITÉRIÉE (PRÉDICTION DU RENDEMENT AU TRAVAIL) ET DES COÛTS D'UTILISATION

Méthode de sélection	Validité critériée approximative*	Coûts d'utilisation
Échantillons de travail	0,54	Modérés à élevés
Tests d'aptitude mentale générale	0,51	Faibles
Entrevues structurées	0,51	Modérés à élevés
Examens de connaissances	0,48	Modérés
Entrevues non structurées	0,38	Faibles à modérés
Centres d'évaluation	0,37	Très élevés
Données biographiques	0,35	Modérés à élevés
Inventaires de personnalité (conscience professionnelle)	0,31	Faibles
Vérification de références	0,26	Faibles
Expérience (nombre d'années)	0,18	Faibles
Inventaires d'intérêts	0,10	Faibles

* Estimation de la validité moyenne proposée par Schmidt et Hunter (1998).

recherches[3]. L'indice de validité utilisé est le coefficient de validité critériée ; il provient de la corrélation entre les résultats de divers échantillons de personnes obtenus à chaque type d'instrument de mesure et leur rendement global au travail.

Malgré leurs limites, les valeurs accordées à ces deux paramètres donnent lieu à une appréciation plus nuancée. Classées en fonction du coefficient de validité seulement (par ordre décroissant), les

3. Les indices de validité sont génériques et n'autorisent pas de distinction selon la méthode de collecte de données, la mesure du rendement utilisé, la nature et le niveau d'emploi considéré, etc. Par exemple, l'estimation de la validité des tests d'aptitude mentale générale varie de 0,23 pour des emplois non spécialisés à 0,58 pour des emplois très complexes de professionnels et de cadres (voir Schmidt et Hunter, 1998). La validité d'une entrevue peut osciller entre 0,20 et 0,57 à mesure que son niveau de structure (standardisation) augmente (Huffcutt et Arthur (1994).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

méthodes qui apparaissent les plus efficaces sont les échantillons de travail, les tests d'aptitude mentale générale, les entrevues structurées et les examens de connaissances. Cependant, si l'on ajoute les coûts d'utilisation, les tests d'aptitude viennent en tête de liste. Les échantillons de travail, malgré leur validité élevée, entraînent des coûts de développement non négligeables selon la complexité de la démarche empruntée. Les entrevues structurées sont également assombries par des coûts qui peuvent être élevés, particulièrement si elles sont conduites en comité : temps de préparation, conduite des entrevues proprement dites, analyse des informations et prise de décision. Quant aux examens de connaissances, avec leur validité substantielle et leurs coûts raisonnables, ils constituent des outils à considérer.

Le cas du centre d'évaluation est intéressant. Son évaluation générale est plutôt faible lorsqu'on tient compte de sa validité moyenne et de ses coûts élevés. Or, une des caractéristiques du centre d'évaluation est de recourir systématiquement aux tests « *In-Basket*[4] » et aux diverses simulations (p. ex., discussion de groupe, jeu de rôle, analyse de cas, etc.). Comment alors expliquer la popularité dont jouissent ces instruments auprès de la plupart des firmes de consultants et de plusieurs agences gouvernementales ? D'autres facteurs sont probablement à l'origine de cet état de fait, comme la validité apparente.

Validité apparente. En plus de la validité et des coûts directs d'utilisation, Melanson et Fontaine (1993) retiennent la validité apparente comme critère d'évaluation. La validité apparente (*face validity*) est la perception qu'un profane peut avoir d'un instrument de sélection. Cette qualité n'a rien à voir avec la validité réelle de l'instrument ou sa rigueur : elle ne porte que sur l'apparence. La validité apparente est élevée lorsqu'une personne, après avoir examiné superficiellement le contenu d'un instrument de mesure, pense qu'il mesure vraiment ce qu'elle croit qu'il devrait mesurer (Catano *et al.*, 1997). Supposons une entrevue au cours de laquelle on pose des questions de mises en

4. Le « *In-Basket* », habituellement traduit par « épreuve du courrier », est une simulation écrite qui a pour but de mesurer les habiletés de gestion telles que la planification, l'organisation, le suivi, la prise de décision ou la délégation.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

situation basées sur des problèmes réels survenus dans le cadre de l'emploi visé ; cette entrevue paraîtra valide aux yeux des candidats et même à ceux des décideurs. À l'opposé, il semblera peu pertinent de faire passer un test d'aptitude mentale composé d'images à compléter ou de mots pour lesquels il faut trouver des synonymes à des candidats à un poste d'ouvrier spécialisé. Pourtant, ce test d'aptitude peut avoir un bon coefficient de validité critériée parce qu'il mesure l'intelligence générale, alors que l'entrevue situationnelle pourrait ne pas être très valide si, par exemple, les grilles de notation ne sont pas remplies par les interviewers de façon juste et systématique.

La validité apparente est particulièrement importante en sélection à cause de son influence marquée sur l'attitude des personnes à l'égard des instruments de mesure. Des instruments dont la validité apparente est élevée ont plus de chances de susciter l'intérêt des **candidats** à s'y soumettre, de projeter l'image d'une évaluation objective et pertinente, ce qui normalement amène les candidats à accepter plus facilement un refus ou une évaluation négative et diminue d'autant leur propension à contester la décision devant une instance légale. Ce dernier aspect est loin d'être négligeable pour l'organisation qui devra défendre en justice sa procédure de sélection, occasionnant une dépense d'énergie et d'argent parfois extravagante, sans pour autant être assurée d'un gain de cause. La validité apparente peut permettre d'éviter un tel litige, comme l'illustre le **cas réel** suivant.

Lors d'une restructuration majeure dans une usine, la direction doit choisir 12 superviseurs parmi le personnel syndiqué. Même si la convention collective n'inclut pas ces nouveaux postes de superviseur, le syndicat demande à ce que l'ancienneté des candidats soit un critère de nomination. Ne pouvant s'entendre avec l'organisation sur la pondération accordée à ce critère, le syndicat augmente la pression et menace de contester en arbitrage le processus de nomination et les outils d'évaluation, ce qui a pour effet de mettre l'organisation sur la défensive. Après mûre réflexion, la direction considère qu'elle ne peut se rendre à la demande syndicale. Elle décide d'aller de l'avant avec le processus de nomination, axé principalement sur les compétences recherchées. Cependant, l'organisation veut à tout prix conserver un climat de travail empreint de collaboration avec ses employés et maintenir son image de bon citoyen corporatif auprès des membres

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

de la petite localité où son usine est implantée depuis plusieurs décennies. À titre d'employeur principal, l'organisation est déterminée à éviter un conflit déchirant qui pourrait opposer des voisins, voire des membres d'une même famille. Dans ces conditions, le processus de nomination doit être juste et équitable, et **perçu comme tel** de la part des candidats et du syndicat. En d'autres mots, les instruments de mesure doivent non seulement être valides, mais aussi avoir un degré élevé de validité apparente.

Avec l'aide d'un spécialiste externe, le service des ressources humaines met sur pied un centre d'évaluation complet, comprenant entre autres une simulation de travail en équipe et un « *In-Basket* » construit sur mesure pour refléter la nouvelle réalité de l'usine. Les candidats sont informés des outils d'évaluation qui seront employés et peuvent même recevoir de l'aide pour leur préparation. La rigueur doit être perceptible, du début à la fin du processus. En fin de compte, l'organisation a gagné son pari : aucun candidat n'a porté plainte et le syndicat n'a pas mis sa menace à exécution. Est-ce par dépit ou par bonne foi, l'histoire ne le dit pas. Malgré ses coûts très élevés, le centre d'évaluation s'est révélé un choix judicieux et certainement rentable. Des tests psychométriques, même valides et abordables, auraient probablement été plus coûteux et certainement moins efficaces pour atteindre les objectifs que l'organisation s'était fixés ou éviter les conséquences qu'elle appréhendait.

Le préjugé favorable suscité par la validité apparente ne touche pas que les candidats ; il contribue aussi à accroître la confiance du **décideur** ou de la personne qui fera usage des résultats. Voilà sans doute ce qui explique en partie la grande popularité des mises en situation, dont fait partie l'épreuve du courrier (*In-Basket*). La validité apparente n'est probablement pas non plus étrangère au fait souvent observé que les membres d'un comité de sélection tendent à accorder plus d'importance aux résultats provenant d'outils comme les simulations ou l'entrevue, qu'à ceux recueillis par des outils moins favorisés par leur apparence, comme les tests psychométriques ou les inventaires de personnalité. Enfin, les **magistrats** ne sont pas insensibles non plus aux charmes de la validité apparente. Lors de poursuites devant les tribunaux, les personnes qui doivent trancher un litige peuvent être influencées par leur propre perception, bien que

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

subjective, de la validité. S'il y a validité apparente, il sera plus aisé de les convaincre de la validité réelle et, partant, du bien-fondé des instruments de mesure employés[5].

Implications légales. Dans plusieurs pays occidentaux, les pratiques de gestion des ressources humaines sont encadrées par diverses lois et réglementations. Au Québec, il y a la *Charte canadienne des droits et libertés*, la *Charte des droits et libertés de la personne* du Québec, le *Code du travail*, les contrats collectifs de travail et les nombreuses politiques gouvernementales en matière d'emploi. Peu importe le cadre législatif visé, si un instrument de mesure fait l'objet d'une contestation, il faudra répondre aux allégations des appelants. Ces considérations légales constituent un autre critère important dans l'évaluation globale d'un instrument de mesure. Cependant, si aucun litige n'est en vue ou si sa probabilité est faible, le critère perd son importance.

Une contestation légale est **possible** seulement si un cadre législatif peut s'appliquer. Par exemple, un employé syndiqué pourra contester une nomination si sa convention collective prévoit des mécanismes clairs et explicites à cet égard (p. ex., une clause qui établit que le poste « *sera pourvu par le salarié qui a le plus d'ancienneté parmi ceux qui ont posé leur candidature, à la condition qu'il puisse satisfaire aux exigences normales de la tâche* »). Quant à la **probabilité** d'une telle contestation, elle dépend principalement de l'attitude du

5. Cette tendance est indirectement soutenue par des chercheurs ayant démontré qu'une défense axée sur des indices de validité critériée, mettant en cause des instruments dont la validité apparente est souvent faible, avait beaucoup moins de chances de convaincre la cour qu'une preuve appuyée sur la validité de contenu, où la validité apparente des instruments est généralement plus élevée (Kleiman et Faley, 1985). En effet, la validité apparente est fréquemment confondue avec la validité de contenu. Sur le plan théorique, cette confusion est une erreur, car les deux concepts sont différents ; la validité apparente est superficielle et subjective alors que la validité de contenu est obtenue au terme d'une démarche systématique et rigoureuse (voir les chapitres 3 et 7). Sur le plan pratique, par contre, il existe une certaine relation entre ces deux notions. Il est exact qu'une validité apparente élevée n'implique pas nécessairement une validité de contenu élevée. Cependant, la réciproque est souvent exacte. Les instruments développés dans un contexte de gestion des ressources humaines et suivant une procédure de validation basée sur le contenu, du genre de celle présentée plus loin dans cet ouvrage, ont aussi une validité apparente élevée dans la plupart des cas. Cela explique que les deux concepts se confondent souvent dans la réalité.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

candidat ou de son représentant. Cette attitude peut être conditionnée par une foule de facteurs comme la validité apparente, le professionnalisme des responsables du processus de sélection ou de promotion, la perception d'équité du candidat à l'égard du processus, le climat de travail ou le militantisme syndical, sans oublier un aspect important, à savoir l'accès par le plaignant à des ressources matérielles, financières et d'expertise. Sans ressources tangibles, préparer un dossier qui résistera à l'offensive de la partie adverse n'est pas une mince affaire.

Bien que chaque cause soit particulière (p. ex., loi ou règlement invoqué, allégations, fardeau de preuve), certains instruments sont plus faciles à défendre que d'autres en raison de leurs qualités métriques propres ; ces qualités découlent principalement de la fidélité et la validité sous toutes leurs formes. Ainsi, tous les outils objectifs avec grille de correction standardisée (tests psychométriques, inventaires de personnalité, examens de connaissances avec clé de correction exhaustive ou entrevues structurées avec grille de notation) ont tendance à présenter une bonne **fidélité**, ce qui est l'un des éléments marquants pour démontrer que le traitement des candidats est équitable et exempt d'erreurs accidentelles (la notion de fidélité fait l'objet d'un chapitre). Il en est de même lorsque les instruments sont construits pour refléter les tâches à exécuter, conformément à la logique de la **validité de contenu** appliquée en gestion des ressources humaines (examens de connaissances, échantillons de travail, simulations ou entrevues structurées fondées sur le poste à combler, centre d'évaluation ou épreuve du courrier). **Validité apparente** aidant pour la plupart de ces instruments, il est dès lors plus facile d'en établir la pertinence et le bien-fondé par rapport à l'emploi en cause (un chapitre porte sur la démonstration de la validité en se basant sur le contenu de l'instrument et deux autres portent sur l'élaboration d'instruments de mesure). Enfin, produire un coefficient de **validité critériée** est une autre façon de démontrer la pertinence d'un instrument (évaluer la validité en se basant sur la relation avec d'autres variables fait l'objet d'un chapitre). Cependant, si l'instrument ne présente pas, en plus, une validité de contenu évidente par rapport au poste, ou au moins une validité apparente, les données sur la validité critériée risquent de ne pas être suffisamment persuasives (revoir les travaux de Kleiman et Faley, 1985) ; c'est souvent le cas avec les tests psychométriques et autres inventaires de personnalité.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Ces derniers outils sont affligés d'un autre facteur de risque du point de vue légal : ils prêtent flanc à des accusations de **discrimination indirecte** en vertu des lois visant à assurer l'équité en emploi de tous les groupes composants la société (Cascio et Sweet, 1989). Par exemple, une population de candidats défavorisés peuvent avoir des résultats systématiquement plus faibles que la population générale à des tests d'aptitudes verbales. De la même manière, l'origine culturelle ou le milieu familial peut avoir un impact sur les scores à un inventaire de personnalité. En soi, de telles différences dans les résultats (appelés « *adverse impacts* » ou « impacts négatifs ») ne prouvent pas qu'il y a effectivement discrimination. Cependant, elles peuvent être suffisantes pour qu'une plainte soit entendue par le tribunal. L'utilisateur de ces tests devra alors démontrer qu'il n'y a pas discrimination dans les faits ou que la discrimination est justifiée ; dans les deux cas, la preuve risque d'être compliquée à faire.

Par ailleurs, deux méthodes de sélection sont à signaler pour leur propension à avoir peu d'impacts négatifs chez les groupes minoritaires ; il s'agit des échantillons de travail et du centre d'évaluation (Cascio et Sweet, 1989). Pour ce qui est du centre d'évaluation, Hoffman et Thornton (1997) reconnaissent qu'il est moins valide et coûte environ 10 fois plus cher que les tests d'aptitudes. Malgré ces désavantages, ils estiment que son utilisation peut être plus rentable à long terme aux États-Unis parce qu'il a moins d'impacts négatifs.

SYNTHÈSE

La validité, les coûts d'utilisation, la validité apparente et les implications légales constituent quatre critères importants susceptibles d'aider à mieux saisir la valeur d'un outil de sélection utilisé dans une situation donnée. En guise de synthèse, les diverses méthodes de sélection apparaissant au tableau 1.1 ont été analysées de nouveau en tenant compte de tous ces critères. Les résultats de cette analyse sont présentés au tableau 1.2. Les valeurs attribuées, il convient de le rappeler, sont approximatives. Ces nouvelles données sont de nature à modifier l'appréciation établie plus tôt de certaines méthodes de sélection. Prenons l'exemple des tests d'aptitudes qui figuraient en tête de liste lorsque seuls les coefficients de validité critériée et les coûts d'utilisation étaient considérés. Ils font piètre figure au chapitre

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 1.2
ANALYSE COMPARATIVE DE DIVERSES MÉTHODES DE SÉLECTION DU PERSONNEL POUR QUATRE CRITÈRES D'ÉVALUATION

Méthode de sélection	Validité critériée approximative	Coûts d'utilisation	Validité apparente	Facilité à être défendue légalement
Échantillons de travail	0,54	Modérés à élevés	Très élevée	Très élevée
Tests d'aptitude mentale générale	0,51	Faibles	Faible	Faible à modérée
Entrevues structurées	0,51	Modérés à élevés	Élevée à très élevée	Élevée à très élevée
Examens de connaissances	0,48	Modérés	Très élevée	Très élevée
Entrevues non structurées	0,38	Faibles à modérés	Faible à modérée	Faible à modérée
Centres d'évaluation	0,37	Très élevés	Élevée à très élevée	Élevée à très élevée
Données biographiques	0,35	Modérés à élevés	Variable	Ne sais pas
Inventaires de personnalité (conscience professionnelle)	0,31	Faibles	Faible à modérée	Faible à modérée
Vérification de références	0,26	Faibles	Faible à modérée	Ne sais pas
Expérience (nombre d'années)	0,18	Faibles	Élevée	Ne sais pas
Inventaires d'intérêts	0,10	Faibles	Modérée à élevée	Ne sais pas

de la validité apparente et la défense en cas de litige. En conséquence, si une organisation veut éviter les inconvénients de la validité apparente et si la possibilité d'un litige est élevée, l'emploi de cet outil de sélection est à déconseiller. Cette nouvelle perspective est cependant très favorable à des méthodes comme l'entrevue structurée, l'examen de connaissances, le centre d'évaluation ou l'échantillon de travail. Dans un contexte de promotion interne régie par une convention collective où les litiges sont hautement probables, les coûts directs d'utilisation auront tôt fait d'être compensés par les économies de frais juridiques, sans compter la meilleure disposition qu'auront les

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

employés à l'égard de la procédure de nomination. Il en sera ainsi également dans une situation de sélection où les candidats proviennent de divers groupes sensibles protégés par une charte des droits.

Des critères génériques d'évaluation. L'évaluation globale d'un instrument de sélection ne peut s'effectuer de manière absolue. Au contraire, la valeur d'un outil est relative à la situation considérée, alors que la pertinence de chacun des critères dépend des objectifs et des préoccupations de l'utilisateur. Idéalement, il revient à chacun de dresser sa propre liste de critères d'évaluation en fonction de ses besoins particuliers. Néanmoins, les quatre critères proposés sont tout de même assez généraux ; de plus, certains d'entre eux renferment à leur tour d'autres critères. Par exemple, les coûts d'utilisation englobent tous les aspects pratiques ayant des répercussions sur les ressources requises à l'application d'un outil de sélection. La facilité à défendre cet outil sur le plan légal est principalement reliée aux diverses formes de fidélité et de validité, incluant la validité apparente[6].

Objet mesuré. Au-delà de tous ces critères d'évaluation, il ne faudrait pas oublier le plus important : la caractéristique ou l'objet mesuré par l'instrument[7]. En effet, tous les instruments de sélection ne mesurent pas avec la même facilité chacune des caractéristiques humaines (voir tableau 1.3). Par exemple, les tests psychométriques sont imbattables pour mesurer les aptitudes et la personnalité, mais réussissent moins bien à saisir les dimensions de la motivation[8]. Les examens de connaissances et les échantillons de travail sont conçus d'ordinaire pour évaluer des connaissances et des habiletés spécifiques à un emploi ou à une famille d'emplois. L'épreuve du courrier et les autres simulations qui se trouvent dans le centre d'évaluation typique pour gestionnaires apprécient une panoplie d'habiletés, le plus souvent de nature cognitive et interpersonnelle. L'entrevue, outil peut-être le plus polyvalent, permet d'évaluer des connaissances, toute la

6. Le coefficient de validité critériée est un indice qui englobe non seulement les objectifs de rendement de l'organisation, mais aussi des aspects plus techniques comme la fidélité et la validité intrinsèque (ou psychométrique) de l'instrument de mesure (voir les chapitres 2 et 4).
7. Campbell (1990) souligne la fâcheuse habitude qu'ont les chercheurs d'organiser leurs résultats en fonction de la méthode de mesure utilisée (tests, entrevue, examens, etc.), laissant dans l'ombre l'objet mesuré par ces méthodes.
8. Les diverses caractéristiques humaines sont définies au chapitre 5, à la section « Spécification du domaine ou du sous-domaine à mesurer ».

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 1.3
EFFICACITÉ APPROXIMATIVE DES MÉTHODES DE SÉLECTION POUR DIVERS OBJETS MESURÉS

Méthode de sélection	Aptitudes	Habiletés et connaissances	Personnalité et autres qualités personnelles	Motivation
Tests d'aptitudes	••••	–	–	–
Inventaires de personnalité	–	–	••••	••
Examens de connaissances	•	••••	–	–
Échantillons de travail	•	••••	•	–
Épreuve du courrier	•	••••	Ne sais pas	–
Autres simulations	•	••••	•	–
Entrevues	••	••••	•••	•••
Expérience	Ne sais pas	••	•	•
Vérification de références	Ne sais pas	••	•	•

Note : Évaluation subjective de l'auteur allant de « totalement inefficace » (–) à « très efficace » (••••).

gamme des habiletés (à l'exception des habiletés psychomotrices) ainsi que les dimensions de la personnalité et de la motivation. L'expérience et la vérification des références sont des outils, comme l'entrevue, qui peuvent être utilisés à diverses fins.

CALCUL DU RETOUR SUR L'INVESTISSEMENT OU ESTIMATION DE L'UTILITÉ ÉCONOMIQUE

Une seconde approche pour établir la valeur globale d'un instrument de mesure consiste à effectuer une véritable analyse de rentabilité, basée sur la différence entre les coûts et les bénéfices. Le ratio ainsi calculé est appelé **utilité** ; il fait référence au gain économique estimé résultant de l'application d'une méthode ou d'un programme. Ce concept, naturel pour les dirigeants habitués à calculer la rentabilité d'une décision ou d'un projet, est tout à fait approprié pour les outils

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

de sélection. Il est alors relativement aisé pour un gestionnaire d'estimer les **coûts** entraînés par l'application ou l'introduction d'un outil de sélection : achat ou développement de l'instrument, nombre de personnes affectées et durée, lors de son application auprès des candidats, de la notation des réponses, de l'analyse et de la prise de décision, utilisation de locaux, frais de transports, etc. Quant aux **bénéfices**, il s'évalue généralement par l'augmentation du rendement entraînée par la méthode de sélection. Ainsi, une méthode de sélection est « utile » économiquement parlant si elle occasionne un accroissement du rendement qui dépasse ses coûts d'administration.

Malgré sa logique des plus élémentaires, le calcul de l'utilité est une procédure rarissime dans la pratique. Lorsque vient le temps de faire les calculs, force est de constater qu'il y a un hic. De prime abord, estimer précisément les bénéfices (soit l'amélioration du rendement) découlant de l'application d'un outil de sélection n'est pas évident dans une situation réelle. Plusieurs auteurs se sont penchés sur ce problème relativement aux outils de sélection et proposent des solutions.

AUGMENTATION DU POURCENTAGE D'EMPLOYÉS SATISFAISANTS (TAYLOR-RUSSEL)

Taylor et Russel (1939, cité dans McCormick et Tiffin, 1974) ont développé une approche pour calculer l'efficacité d'une méthode, d'un outil de sélection. Simple d'emploi, leur approche nécessite de connaître la valeur de trois paramètres. Le premier est la **validité**, exprimée par le coefficient de corrélation entre les résultats obtenus à cette méthode de sélection et une mesure du rendement au travail (coefficient de validité critériée). Le deuxième est le **taux de sélection**, défini par la proportion de candidats qui ont été sélectionnés par rapport au nombre total de candidats évalués par cet outil. Par exemple, si 100 personnes se présentent à un examen de connaissances et que 15 d'entre elles sont choisies pour occuper un poste, alors le taux de sélection est de 15 % pour cet examen. Le troisième est le **pourcentage de candidats susceptibles de convenir au poste** parmi la population des candidats qui se soumettent à l'outil de sélection. C'est un indice du degré de qualification de la main-d'œuvre par rapport au poste à pourvoir ou, inversement, un indice du degré de difficulté du poste. Dans le cas d'un emploi difficile, par

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

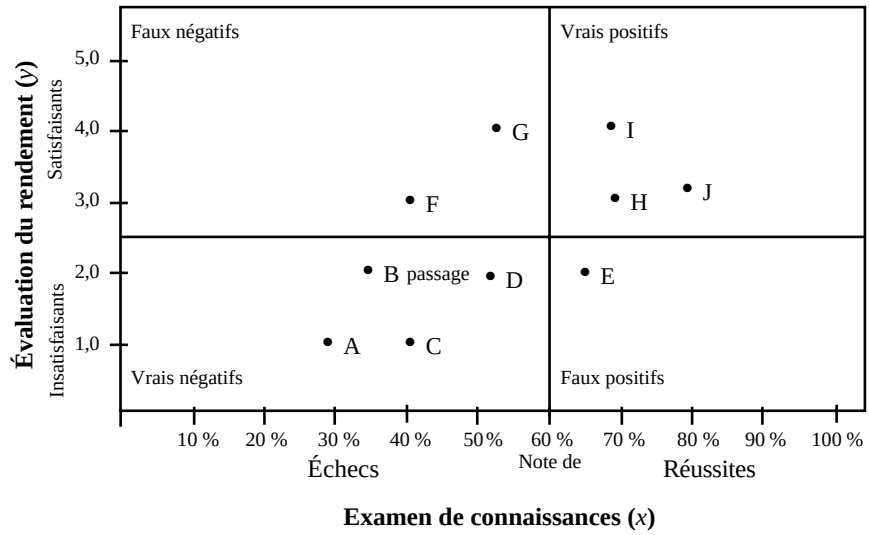
exemple, ce pourcentage pourrait être de 25 %, indiquant que seul le quart des candidats parmi tous ceux évalués seraient capables de faire le travail de manière satisfaisante.

Exemple. Prenons une organisation fictive qui désire évaluer l'efficacité d'un examen de connaissances comme moyen de sélection. Cet examen, qui s'ajoute au processus de sélection en vigueur, est administré à un échantillon de 10 employés déjà en poste. La note de passage est fixée à 60 % (voir figure 1.1). En guise de critère, le rendement de ces employés est évalué par leur superviseur au moyen d'une échelle en cinq points. Un employé qui obtient une note d'évaluation inférieure à 2,5 n'est pas considéré comme répondant pleinement aux exigences du poste. Les données de la figure 1.1, présumées représentatives de l'ensemble des employés de l'organisation, permettent d'analyser ce qui se passerait si cet examen était utilisé comme outil de sélection. Les candidats sélectionnés grâce à ce nouveau test sont appelés « positifs », par opposition à ceux rejetés et désignés comme « négatifs ». Les « vrais positifs » sont ceux dont le rendement est satisfaisant et qui ont, à juste titre, été sélectionnés. En revanche, ceux dont le rendement est insatisfaisant mais qui ont été choisis quand même, par erreur, sont qualifiés de « faux positifs ». De la même manière, les « faux négatifs » sont des candidats satisfaisants mais qui ont été rejetés à tort, alors que les « vrais négatifs » sont de véritables candidats insatisfaisants qui n'ont pas été sélectionnés. La corrélation entre les résultats à l'examen et le critère de rendement a été estimée à 0,60.

Est-ce que ce nouvel outil de sélection peut être valable pour l'organisation ? La réponse est oui. L'application de cet examen de connaissances comme outil de sélection permettrait de faire passer le pourcentage d'employés en poste jugés satisfaisants de 50 % à 75 %. Le raisonnement est fort simple. Les données de la figure 1.1 indiquent que, présentement, 5 employés sur 10 (soit F, G, H, I et J) ont un rendement évalué à plus de 2,5, ce qui traduit un pourcentage d'employés jugés satisfaisants de 50 %. Si le test de connaissances était appliqué comme moyen de sélection avec une note de passage de 60 %, seulement quatre candidats seraient sélectionnés (les « positifs » E, H, I et J). Cependant, trois d'entre eux (les « vrais positifs » H, I et J) auraient un rendement satisfaisant, soit 75 % de tous les candidats maintenant sélectionnés grâce à l'introduction de cet examen de connaissances.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 1.1
DIAGRAMME DE DISPERSION ENTRE LES RÉSULTATS À UN TEST DE CONNAISSANCES ET LE RENDEMENT AU TRAVAIL POUR UN ÉCHANTILLON FICTIF DE 10 EMPLOYÉS



Tables de Taylor-Russel. Des tables compilées par Taylor et Russel (1939, cité dans McCormick et Tiffin, 1974) permettent de déterminer directement le pourcentage d'employés sélectionnés qui seront jugés satisfaisants, en fonction des diverses combinaisons de validité, de taux de sélection et de pourcentage de candidats susceptibles de convenir au poste. Un extrait de ces tables apparaît au tableau 1.4 et leur utilisation est illustrée par l'exemple de la figure 1.1. Premièrement, il faut choisir la portion du tableau dont le « **pourcentage de candidats susceptibles de convenir** » correspond aux données de la situation envisagée. Dans l'exemple, ce pourcentage est de 50 %. Il faut donc se reporter à la troisième section du tableau intitulé « Pourcentage d'employés qui seront jugés satisfaisants, si 50 % des candidats sont susceptibles de convenir ». Deuxièmement, il faut localiser, dans la colonne de gauche, la valeur appropriée du **coefficient de validité**, établie à 0,60 dans l'exemple. Elle se trouve à la cinquième ligne. Troisièmement, il est nécessaire

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 1.4
EXTRAITS DES TABLES DE TAYLOR-RUSSEL

	Taux de sélection			
Validité critériée	**0,2**	**0,4**	**0,6**	**0,8**
Pourcentage d'employés qui seront jugés satisfaisants, si **30 %** des candidats sont susceptibles de convenir.				
0,20	40 %	37 %	34 %	32 %
0,30	46 %	40 %	37 %	33 %
0,40	51 %	44 %	39 %	34 %
0,50	58 %	48 %	41 %	35 %
0,60	64 %	52 %	43 %	36 %
0,70	72 %	57 %	46 %	37 %
Pourcentage d'employés qui seront jugés satisfaisants, si **40 %** des candidats sont susceptibles de convenir.				
0,20	51 %	48 %	45 %	43 %
0,30	57 %	51 %	47 %	44 %
0,40	63 %	56 %	50 %	45 %
0,50	69 %	60 %	53 %	46 %
0,60	75 %	64 %	55 %	48 %
0,70	82 %	69 %	58 %	49 %
Pourcentage d'employés qui seront jugés satisfaisants, si **50 %** des candidats sont susceptibles de convenir.				
0,20	61 %	58 %	55 %	53 %
0,30	67 %	62 %	58 %	54 %
0,40	73 %	66 %	61 %	56 %
0,50	78 %	70 %	63 %	57 %
0,60	84 %	75 %	66 %	59 %
0,70	90 %	80 %	70 %	60 %
Pourcentage d'employés qui seront jugés satisfaisants, si **60 %** des candidats sont susceptibles de convenir.				
0,20	71 %	67 %	65 %	63 %
0,30	76 %	71 %	68 %	64 %
0,40	81 %	75 %	70 %	66 %
0,50	86 %	79 %	73 %	67 %
0,60	90 %	83 %	76 %	69 %
0,70	94 %	87 %	80 %	71 %
Pourcentage d'employés qui seront jugés satisfaisants, si **70 %** des candidats sont susceptibles de convenir.				
0,20	79 %	77 %	75 %	73 %
0,30	84 %	80 %	77 %	74 %
0,40	88 %	83 %	79 %	75 %
0,50	91 %	87 %	82 %	77 %
0,60	95 %	90 %	85 %	79 %
0,70	97 %	93 %	88 %	80 %

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

de se déplacer sur cette ligne vers la droite jusqu'à la colonne où le taux de sélection, indiqué en haut du tableau, correspond à celui de l'exemple ; pour un **taux de sélection** de 0,40, c'est la deuxième colonne. Il ne reste plus qu'à lire le chiffre, 75 %, qui se trouve à la croisée de ces deux alignements : c'est le pourcentage prévu d'employés jugés satisfaisants si un test était utilisé dans une situation caractérisée par ces trois paramètres. En d'autres termes, le pourcentage initial de 50 % serait passé à 75 %.

Situations qui augmentent l'utilité d'un instrument de sélection. Au-delà des problèmes d'application dénoncés plus loin, réels et sans doute insurmontables dans plusieurs cas, les tables de Taylor-Russel permettent aux gestionnaires de savoir quand un instrument est susceptible de donner les meilleurs résultats (Schneider et Schmitt, 1986). Premièrement, le pourcentage d'employés satisfaisants sera toujours assez élevé si le **taux de sélection est faible**. Par exemple, pour un taux de sélection de 0,20, cette proportion se situe systématiquement entre 50 % et 97 %, à quelques exceptions près, ce qui souligne l'importance d'évaluer un nombre de candidats beaucoup plus élevé que le nombre de postes à combler et la nécessité de recourir à des sources de recrutement efficaces. Il faut toutefois signaler qu'une telle approche occasionnera des coûts additionnels, tant pour le recrutement que pour l'application des instruments de sélection à un plus grand nombre de personnes.

Deuxièmement, même avec une validité plutôt faible, un outil peut donner des résultats intéressants si le taux de sélection est faible. Par exemple, avec une validité de 0,30 et un taux de sélection de 0,20, un test peut donner lieu à un pourcentage d'employés satisfaisants de 46 % lorsqu'il est appliqué à des candidats dont 30 % sont susceptibles de convenir au poste. Ces proportions passent respectivement à 57 %, 67 %, 76 % et même 84 % pour des candidats potentiellement satisfaisants à 40 %, 50 %, 60 % et 70 %. Autrement dit, un test dont la validité est faible peut tout de même être efficace si le **nombre de candidatures est élevé**.

Troisièmement, un outil de sélection atteint son maximum au regard de l'augmentation relative de la proportion d'employés satisfaisants lorsqu'il y a **peu de candidats susceptibles de convenir au poste**. Par exemple, avec une validité de 0,40 et un taux de sélection

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

de 0,20, un test donnera un pourcentage de 51 % d'employés satisfaisants si 30 % des candidats sont potentiellement acceptables, ce qui est une amélioration relative de 70 % (51 % – 30 % = 21 %, puis 21 % ÷ 30 % = 70 %). En comparaison avec une situation où 70 % des candidats sont susceptibles d'être satisfaisants, l'amélioration serait à peine de 26 % (88 % – 70 % = 18 %, puis 18 % ÷ 70 % = 26 %). Un outil de sélection a donc plus de chances d'être rentable si la proportion de la main-d'œuvre qualifiée est faible. Autrement dit, il ne peut améliorer une main-d'œuvre qui serait déjà largement satisfaisante.

Voilà trois principes qu'un gestionnaire devrait considérer avant de porter un jugement sur un instrument de sélection. Et il n'est pas nécessaire de se livrer à des calculs compliqués pour voir si une situation est propice à l'usage d'un outil de sélection ou pour trouver des moyens de rendre une situation plus favorable.

Joindre ou remplacer une procédure existante. Évaluer l'efficacité d'un outil de sélection implique de comparer deux situations : on compare la situation dans laquelle l'outil est employé à une autre dans laquelle il ne l'est pas. Pour ce faire, deux scénarios différents peuvent être envisagés. Dans les exemples mentionnés plus haut, on a considéré un scénario dans lequel l'outil analysé s'ajoutait au processus de sélection en vigueur ; on a voulu connaître l'apport additionnel de ce nouvel outil, une fois **ajouté à la procédure ou aux outils existants**. Dès lors, on a comparé la situation mettant en cause le processus de sélection existant à une situation où l'outil analysé serait ajouté à ce processus. Dans ces circonstances, il est justifié d'estimer le pourcentage de candidats susceptibles de convenir au poste simplement en se basant sur les employés déjà en poste dans l'organisation, ces derniers ayant été sélectionnés par la procédure en vigueur. Un autre scénario aurait pu intéresser le gestionnaire : **remplacer la procédure existante** de sélection par une autre (Boudreau, 1991). Les paramètres de ce nouveau cas de figure devraient alors refléter ceux de la population des candidats non sélectionnés par la procédure existante. Les employés en place ne sont pas représentatifs de cette population de candidats et ne peuvent plus servir d'échantillon pour estimer le pourcentage des candidats susceptibles de convenir ; il en est de même pour la validité. Lorsqu'un instrument s'ajoute aux autres, sa validité devrait être estimée auprès des

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

employés déjà en place et lorsqu'un outil en remplace d'autres, il faut recourir à la validité estimée auprès de la population des candidats non sélectionnés.

Deux lacunes. La méthode de Taylor-Russel n'a pas résolu complètement ni parfaitement le problème de l'estimation des bénéfices engendrés par l'utilisation d'un outil de sélection. En effet, déterminer l'amélioration du pourcentage d'employés jugés satisfaisants ne donne pas une estimation complète de l'utilité du nouvel instrument de sélection. Il conviendrait de traduire cette augmentation en **gain économique**, puis d'en soustraire les coûts occasionnés par l'application de l'outil. On s'en doute, la traduction obligée en gain économique n'est pas un calcul aisé à effectuer, ce qui explique probablement pourquoi la démarche est si peu appliquée. L'approche de Taylor-Russel souffre d'une autre lacune, cette fois, d'ordre méthodologique. Les employés sont classés seulement en **deux groupes par rapport au rendement** : ils sont ou bien « satisfaisants », ou bien «insatisfaisants». Une fois classés, tous les employés de la même catégorie, « satisfaisants » ou « insatisfaisants », sont considérés comme ayant un rendement égal, alors que ce n'est pas le cas dans les faits (Boudreau, 1991 ; Schneider et Schmitt, 1986). Simplifier ainsi la réalité peut fausser l'estimation des retombées d'une méthode de sélection.

CALCUL DU GAIN ÉCONOMIQUE (BROGDEN-CRONBACH-GLESER)

Il existe une approche qui prévient les deux inconvénients de la méthode de Taylor-Russel. Cette approche permet d'évaluer l'amélioration qui découle de l'application d'un outil de sélection directement du point de vue des gains économiques, du bénéfice net pour l'organisation. De plus, elle traite le rendement sur une échelle continue, et non pas seulement en fonction de deux valeurs (« satisfaisants/non satisfaisants »). Cette approche est connue sous le nom de ses trois auteurs, soit Brogden-Cronbach-Gleser (Brogden, 1949 ; Cronbach et Gleser, 1965). Suivant cette démarche, le gain économique d'un outil de sélection dépend de cinq paramètres, comme on peut le constater dans l'équation suivante (Boudreau, 1991 ; Cook, 1988 ; Schneider et Schmitt, 1986) :

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Gain économique par employé par année = $(r_{xy} \times S_y \times Z_x) - (C \div TS)$

où r_{xy} : validité de l'outil de sélection exprimée par le coefficient de corrélation entre cet outil et le rendement dans l'emploi considéré.

S_y : écart type du rendement des employés embauchés sans l'usage de l'outil de sélection, exprimé en valeur monétaire.

Z_x : résultat moyen à l'outil de sélection des personnes sélectionnées, exprimé en note standard (Z) basée sur l'ensemble des candidats évalués pour cet emploi.

C : coût pour appliquer l'outil de sélection à un candidat.

TS : taux de sélection.

Exemple. Si l'on connaît la valeur des paramètres, cette formule est simple à appliquer. Supposons une organisation qui doit pourvoir des postes dont le salaire annuel est de 30 000 $ et dont l'écart type du rendement (S_y) des candidats est estimé à 40 % du salaire, soit 12 000 $ (voir chapitre 7). Cette organisation a confié la sélection en sous-traitance à une firme d'experts qui utilise un test d'aptitude dont la validité reliée au critère de rendement (r_{xy}) est de 0,50. Les candidats sélectionnés ont obtenu en général des résultats se situant à un écart type au-dessus de la moyenne des postulants à cet emploi, ce qui fait que Z_x vaut 1,0. La firme demande 300 $ par candidat (C). Il lui faut évaluer 10 candidats pour en trouver 2 recommandables, de sorte que le taux de sélection (TS) est de 0,2. Le gain financier par employé sélectionné est :

$(0{,}50 \times 12\,000\ \$ \times 1) - (300\ \$ \div 0{,}2) = (6\,000\ \$) - (1\,500\ \$) = 4\,500\ \$$

Chaque employé embauché suivant la procédure évaluée de sélection rapporte environ 4 500 $ de plus par année qu'un employé qui aurait été sélectionné au hasard parmi ces mêmes candidats. Pour un gestionnaire, il est intéressant de constater que les gains économiques sont plus élevés lorsque l'écart type du rendement (S_y) des candidats est, lui aussi, élevé. En d'autres mots, il est plus rentable de recourir à un outil de sélection lorsque les candidats diffèrent beaucoup entre eux par rapport à leur capacité d'occuper le poste convoité (cette idée est traitée dans le chapitre 3, à la section « Rationnel de la sélection du personnel »).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Remarques. Comme pour l'approche de Taylor-Russel, la **population de candidats** qui servira à estimer les valeurs des divers paramètres doit être appropriée au scénario qui intéresse l'organisation (Boudreau, 1991). S'il s'agit d'un ajout d'un outil à une procédure de sélection déjà en vigueur, l'écart type du rendement (S_y) et la validité (r_{xy}) doivent être évalués auprès des employés en place, déjà sélectionnés. S'il s'agit d'un remplacement d'une méthode de sélection, alors il faut reprendre les estimations en fonction d'une population de candidats non sélectionnés. Finalement, l'application de la formule Brogden-Cronbach-Gleser repose sur **deux prémisses**. Premièrement, le prédicteur et le critère sont des variables continues ; ils ont des distributions identiques et la relation qui les unit est linéaire (Schneider et Schmitt, 1986). Deuxièmement, lorsque la sélection est effectuée, les candidats sont classés en fonction de leur résultat au nouvel outil (prédicteur), puis embauchés en commençant par le plus fort jusqu'à ce que le nombre désiré de candidats soit obtenu (Boudreau, 1991) ; en anglais, cette façon de procéder s'appelle un *top down*. Pour que cette dernière prémisse soit possible, il faut que tous les candidats choisis acceptent l'offre d'emploi (Cascio, 1993).

A) L'estimation de l'écart type du rendement en dollars : un obstacle en voie d'être surmonté

Pour traduire l'efficacité d'une méthode de sélection en gain économique avec l'approche de Brogden-Cronbach-Gleser, il faut connaître au préalable la valeur du rendement des employés, ou, plus précisément, la valeur économique des différences de rendement entre employés. Estimer la valeur de ce paramètre, c'est-à-dire l'écart type du rendement des employés embauchés sans l'usage de l'outil de sélection (ou S_y), représente l'obstacle le plus difficile à surmonter. Dans les nombreux méandres à travers lesquels nous entraînent les chercheurs, le praticien finira par trouver une solution intéressante.

L'approche comptable. Les premières tentatives ont été faites par des comptables qui se sont attelés à la tâche de calculer la valeur nette de la production d'un employé : le nombre d'unités produites par période de temps, représentant chacune tel montant du volume d'affaires, moins les coûts de cette production, etc. Après avoir effectué ces calculs pour un ensemble d'employés, il ne resterait plus qu'à compiler l'écart type de la distribution obtenue pour connaître S_y. Comme l'ont relaté Boudreau (1991) et Cook (1988), on s'est vite

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

rendu compte de la témérité de l'aventure. Les comptables ne sont pas parvenus, même pour des postes relativement simples de production, à évaluer objectivement la valeur économique de la contribution d'un employé. En plus d'être complexes et coûteuses, leurs démarches demeurent encore aujourd'hui empreintes d'un degré trop élevé d'arbitraire et de subjectivité, surtout pour les postes dont l'unité de rendement et ses conséquences sont difficiles à définir (gestionnaires, professeurs, service à la clientèle, etc.).

Estimation globale par des experts. Pour résoudre ces difficultés, Schmidt *et al.* (1979) proposent de demander à des superviseurs de juger ce que vaut pour l'organisation une différence de rendement de un écart type chez leurs subordonnés. Selon ces auteurs, les supérieurs immédiats sont très bien placés pour observer, dans le quotidien, les variations de rendement de leurs employés. On demande alors à des supérieurs d'expérience d'estimer en dollars la **valeur du travail réalisé** par un employé moyen, c'est-à-dire dont le rendement se situe approximativement au 50e percentile par rapport aux autres employés occupant le même emploi. Ensuite, ils sont amenés à estimer la valeur d'un bon employé, celui qui est environ au 85e percentile (rendement meilleur que 85 % des autres employés), et parfois, celle d'un employé plus faible, qui occuperait le 15e percentile. Si le rendement suit une distribution normale, les percentiles 85e et 15e se situent à un écart type environ de part et d'autre du centre de la distribution, soit le 50e percentile (voir les propriétés de la distribution normale). Par conséquent, la différence de valeur entre le rendement de l'employé moyen et celui dont le rendement est au 85e percentile, ou au 15e percentile, correspond à S_y.

Pour faciliter la tâche des superviseurs, on leur demande d'estimer ce qu'il en aurait coûté à l'organisation si elle avait requis les services d'une firme externe pour fabriquer le même produit ou offrir le même service. Afin de minimiser les biais et les erreurs de mesure propres à chaque superviseur-expert, on a eu recours aux estimations de plusieurs superviseurs. On calcule ensuite la moyenne des estimations pour l'ensemble de ces superviseurs. Dans une étude pilote portant sur des analystes de budgets, Schmidt *et al.* (1979) ont sollicité 62 superviseurs pour qu'ils estiment la valeur de leurs analystes moyens et celle de leurs analystes au 85e percentile. La différence moyenne a été de 11 327 $, ce qui est l'estimation de S_y ou la valeur d'une différence de rendement dont l'ampleur est de un écart type.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Trois lacunes. Si l'approche globale présente l'avantage d'être applicable, elle a néanmoins fait l'objet de plusieurs critiques (Boudreau, 1991 ; Cook, 1988 ; Schuler et Guldin, 1991). Premièrement, les personnes trouvent souvent **difficile d'évaluer** ainsi la valeur monétaire de divers niveaux de rendement. Deuxièmement, la prémisse selon laquelle le rendement se distribue en suivant une **courbe normale** n'est pas prouvée (voir chapitre 7, à la section « Distribution normale »). En fait, il serait prudent de penser que la distribution est spécifique à chacune des situations et, en conséquence, la normalité devrait être vérifiée dans chaque cas. Troisièmement, il peut y avoir beaucoup de **variation** entre les superviseurs-experts en ce qui concerne leurs estimations et leur façon d'y arriver. Il faut dire que la démarche est dangereusement sensible à la subjectivité. En fait, on ne sait pas trop sur quelle base ou suivant quel rationnel chaque superviseur-expert en arrive à ses estimations, de sorte qu'il est difficile de donner une signification exacte à la valeur moyenne obtenue de l'écart type. Pour réduire cette variation, Schuler et Guldin (1991) recommandent entre autres 1) de former davantage les experts sur les différentes facettes de la méthode et 2) de leur donner en feed-back la valeur moyenne que l'ensemble des experts a estimé pour le rendement de l'employé moyen (50^{e} percentile) afin de fournir un point d'ancrage avant leur estimation des autres niveaux de rendement (p. ex., employés au 85^{e} et au 15^{e} percentile). À ce jour, les quelques recherches comparant les valeurs obtenues par la méthode de l'estimation globale à des mesures objectives du rendement ont donné lieu à des résultats encourageants, mais complexes.

Estimation individuelle CREPID. Du nom de ses auteurs (Cascio, 1982 ; Cascio et Ramos, 1986), la méthode CREPID (*Cascio-Ramos Estimate of Performance in Dollars*) est une procédure complexe et détaillée pour estimer la valeur économique de S_y. La **démarche** peut être résumée succinctement de la manière suivante (Boudreau, 1991 ; Schneider et Schmitt, 1986).

1. L'emploi considéré est décomposé en ses activités principales.
2. Chacune des activités est cotée par rapport à deux dimensions : temps/fréquence et importance. Ces deux cotes sont ensuite combinées pour donner une pondération de la valeur totale de chaque activité.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

3. Une valeur monétaire est attribuée à chaque activité en répartissant le salaire annuel au prorata des pondérations obtenues à l'étape précédente.
4. On demande alors à des superviseurs d'évaluer le rendement individuel d'un échantillon d'employés sur chaque activité. L'échelle proposée pour cette tâche s'échelonne de 0 à 2,00, où 1,00 indique un rendement moyen, 2,00 un rendement au 99e percentile, 0,50 un rendement au 25e percentile et ainsi de suite.
5. Pour transformer ces évaluations en dollars, il s'agit de les multiplier par la valeur monétaire attribuée à l'étape 3.
6. Après que chaque activité d'un employé ait reçu ainsi une valeur en dollars, il ne reste qu'à additionner ces valeurs pour obtenir la valeur totale du rendement de cet employé pour une année.
7. Finalement, la moyenne et l'écart type de la valeur totale du rendement pour l'ensemble de l'échantillon d'employés sont calculés. La moyenne devrait correspondre approximativement au salaire annuel alors que l'écart type devient l'estimation de S_y.

Des variations de cette méthode ont été proposées par la suite (voir Boudreau, 1991). Cette nouvelle approche offre l'avantage de rendre plus explicite le processus d'évaluation des employés et d'estimation de la valeur monétaire de leur rendement. Comparée à la méthode précédente de l'estimation globale, l'approche CREPID donne lieu à des estimations de l'écart type de la valeur monétaire du rendement (S_y) plus basses, souvent juste en dessous de 40 % du salaire annuel, comportant moins de variation et reflétant davantage les valeurs obtenues par l'approche comptable (Boudreau, 1991 ; Schuler et Guldin, 1991). Cependant, certains trouvent l'approche CREPID lourde et coûteuse (Schuler et Guldin, 1991).

Règle en proportion du salaire ou des biens et services produits. Cette troisième méthode est des plus pratiques. Comme l'indique Boudreau (1991), l'estimation de S_y est obtenue simplement en multipliant, par une proportion donnée, le salaire annuel ou la valeur moyenne des biens et services produits dans l'emploi concerné. C'est une approche qui a été proposée à la suite d'observations portant sur la valeur relative de S_y en comparaison avec celle du salaire annuel ou de celle des biens et services produits. De cette façon, Hunter et Schmidt (1982) ont estimé, après avoir analysé quelques études empiriques,

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

que S_y correspondait en moyenne à une proportion variant de 40 % à 70 % du salaire annuel. Ils ont également observé que, pour des emplois routiniers, S_y vaut approximativement 20 % de la valeur des biens et services produits, proportion qu'ils estiment à environ 35 % du salaire annuel (Schmidt et Hunter, 1983). Ils ont aussi remarqué que cette proportion, en rapport aux biens et aux services produits, peut augmenter jusqu'à environ 32 % pour des emplois moyennement complexes et à 48 % pour des emplois de professionnels (Hunter, Schmidt et Judiesch, 1990 ; Schmidt *et al.*, 1988). Une règle est suggérée : *l'écart type du rendement exprimé en monnaie (S_y) peut être estimé de façon conservatrice à* ***40 % du salaire annuel*** *ou à* ***20 % des biens et services*** *produits* (Schmidt et Hunter, 1998 ; Schmidt, Hunter, Outerbridge et Trattner, 1986).

Cette règle est-elle fiable ? C'est ce que Boudreau (1991) a cherché à savoir en examinant systématiquement 108 estimations de S_y obtenues dans diverses études. Pour 44 estimations de S_y calculées en proportion des **biens et services** produits, seulement 2 n'atteignent pas les 20 %, 13 sont dans la fourchette de 20 % à 35 % et 29 dépassent 35 %. Par contre, des 64 estimations en rapport au **salaire**, 24 sont inférieures à 40 %, 18 demeurent entre 40 % et 70 % et 22 excèdent la limite de 70 %. En conclusion, utiliser 40 % du salaire annuel aurait tendance à surévaluer S_y alors que 20 % de la valeur des biens et services serait conservateur. Il revient donc au décideur de poser un choix en fonction des conséquences de la décision à prendre. Il convient de noter la grande dispersion de S_y, qui varie notamment selon la complexité de l'emploi. Par conséquent, un gestionnaire qui s'en remet à une proportion générique court le risque de choisir une valeur qui ne correspond pas à la situation particulière qui nous intéresse. Mais, au moins, voilà un barème simple qui lui permet d'appliquer la formule Brogden-Cronbach-Gleser et d'avoir enfin un aperçu réaliste de l'efficacité économique de ses outils de sélection.

Les gains économiques mesurés par trois indicateurs de productivité. Les gains économiques d'une pratique de gestion, que ce soit une méthode de sélection ou toute autre intervention, peuvent être définis par divers indices de productivité. Les trois plus importants jusqu'ici dans la littérature ont été l'augmentation du volume d'affaires, la réduction des coûts et les profits nets (Boudreau, 1991). Par exemple, trois des méthodes présentées pour estimer S_y (estimation globale, CREPID et règles de proportion) portent sur la **valeur**

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

des biens et des services produits : ce qu'il en coûterait si ces biens et ces services étaient fournis par une firme externe ou ce qu'il en coûte par rapport au salaire annuel de l'employé. Cette façon de concevoir les gains de productivité relève clairement de la logique du volume d'affaires. En outre, nous avons relevé dans la littérature plusieurs études recourant au calcul de la réduction des coûts comme mesure d'utilité (p. ex., la réduction des coûts de formation, la réduction du salaire ou du nombre d'employés, la diminution d'achat d'équipement, l'abaissement des accidents, du taux de roulement). Par ailleurs, l'approche de Brogden et de ses collègues (Brogden, 1949 ; Cronbach et Gleser, 1965) misait initialement sur le calcul comptable des revenus moins les coûts, c'est-à-dire sur la notion de **profit**. Boudreau (1991) considère que le mérite de chacune de ces façons de mesurer les retombées financières doit être évalué en fonction du contexte de l'organisation. Si elles reflètent toutes une facette objective de la productivité, elles ont cependant l'inconvénient d'ignorer des facteurs comme le climat de travail, la loyauté des membres de l'organisation, le respect des règles d'éthique et d'autres facettes plus fluides de l'organisation.

B) Raffinements ou le difficile équilibre entre complexité et pragmatisme

Les travaux de Schmidt, Hunter et de leurs collaborateurs publiés en 1979 semblent avoir donné un nouveau souffle à la recherche sur l'utilité économique en psychologie industrielle (Schneider et Schmitt, 1986). Plusieurs méthodes et leurs variantes ont vu le jour, telles que l'estimation globale par des experts, l'approche CREPID ou les règles de proportion. De nombreuses améliorations ont déjà été apportées à ces méthodes, mais au prix d'une complexité accrue.

Complexité croissante. Boudreau propose d'innombrables ajouts et pistes d'amélioration, comme l'intégration de notions d'analyse financière et économique (Boudreau, 1991). Par exemple, il serait nécessaire d'ajuster l'estimation de l'utilité en fonction de trois facteurs supplémentaires, dont l'effet conjugué est substantiel, puisqu'il peut retrancher de 30 % à 80 % des gains estimés. Le premier concerne la **variabilité des coûts** en fonction du rendement d'un employé. En effet, comme un employé plus performant peut engendrer des coûts additionnels (usage de plus de matériel, augmentation des inventaires, budget de déplacement, augmentation de salaire ou prime au

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

rendement, etc.), ces coûts doivent être soustraits du gain économique total. Le deuxième paramètre a trait aux **flux monétaires** qui doivent être actualisés pour tenir compte de l'inflation et des taux d'intérêts. En effet, la plupart des coûts sont engagés lors de l'embauche alors que les bénéfices se concrétiseront plus tard, ce qui fait que 100 $ dépensés aujourd'hui valent plus que les 100 $ qui seront économisés dans deux ans. Le troisième paramètre s'applique aux **taxes et aux impôts** que l'État, « ce partenaire silencieux », prélève sur tout profit additionnel. Ce n'est donc pas le gain économique brut qu'il faut considérer dans le calcul de l'utilité, mais le gain net.

Ces facteurs ne sont pas les seuls dont il faudrait tenir compte. Il y a le nombre d'années pendant lesquelles l'employé sélectionné occupera le poste, étant donné que cette valeur a un effet multiplicateur sur les gains totaux. Il y a la distinction entre les coûts uniques (p. ex., développement et introduction d'une nouvelle méthode de sélection) et les coûts récurrents (p. ex., application de la méthode de sélection à chaque candidat). Il y a aussi des facteurs importants, mais que l'on ne peut estimer, comme le coût des employés qualifiés rejetés par erreur et qui passent chez le concurrent. Bref, en lisant l'imposante synthèse de Boudreau (1991), on se rend compte à quel point il est complexe de mesurer le retour sur l'investissement : il y a tant de coûts et de bénéfices, directs et indirects, tant de facteurs qui entrent dans l'équation.

Réalité organisationnelle. On peut se demander si c'est en raison de sa complexité que le concept d'utilité est si peu appliqué. En effet, combien de praticiens, même avec une formation spécialisée, sont intéressés à approfondir, ou au moins à s'initier aux éléments du calcul de l'utilité ? Voilà pourtant une réalité bien présente dans le monde organisationnel. Il va de soi qu'il faut calculer la rentabilité d'un projet d'acquisition ou d'agrandissement, du lancement d'un nouveau produit, ou même les retombées économiques de la tenue d'un congrès ou d'une campagne visant à attirer les touristes, mais penser à une pratique de gestion des ressources humaines du point de vue coûts-bénéfices, le réflexe n'y est pas. Et a fortiori, si la démarche est compliquée !

Jusqu'où devons-nous aller dans cette voie de la rigueur méthodologique toujours croissante ? Est-ce que toute cette complexité additionnelle en vaut vraiment la peine ? Les moyens mis en œuvre

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

pour prendre une décision ou résoudre un problème doivent être proportionnels à l'importance de la situation et de ses conséquences. En pratique, l'accroissement de la complexité ne peut se justifier que par la valeur de l'information qu'elle permet d'ajouter. Si une approche plus simple peut donner sensiblement les mêmes résultats, alors cela ne vaut pas la peine d'investir plus d'efforts. C'est donc au gestionnaire que revient la responsabilité de déterminer le degré de rigueur.

C) Une démarche qui porte des fruits

Des résultats de recherche formidables. Malgré les critiques et les lourdeurs reprochées, les études sur l'utilité ont eu des retombées inédites. D'abord, les pratiques de sélection du personnel sont devenues enfin **rentables**, et même très rentables. Au terme d'une revue de littérature sur 21 études empiriques visant le calcul de l'utilité pour 48 interventions, Boudreau (1991) en est arrivé à la conclusion que les programmes de sélection sont largement avantageux pour l'organisation ; presque toutes les études font état de gains qui dépassent nettement les coûts. Dans les grandes organisations, les gains se chiffrent souvent en millions de dollars, surtout lorsqu'un nombre élevé de personnes sont touchées par le programme de sélection. Les méthodes de sélection les plus valides, mais aussi les plus coûteuses, produisent le meilleur retour sur l'investissement. Même l'entrevue, dont certaines formes sont peu valides, est profitable. De façon étonnante, les coûts additionnels associés à l'amélioration d'un processus de sélection se sont révélés infimes comparés aux bénéfices, même lorsque ces derniers étaient estimés de façon conservatrice. Il se peut, selon Smith, Gregg et Andrews (1989), que la sélection devienne à l'avenir une des pratiques de gestion du personnel les plus rentables. Il faut alors que les gestionnaires en ressources humaines investissent davantage dans le développement d'un processus de sélection rigoureux. La diffusion de ces résultats de recherche devrait accroître leur sentiment de **confiance** envers leurs propres outils et surtout leur fournir des **arguments** à présenter à la haute direction qui n'a pas toujours tendance à considérer les pratiques de gestion des ressources humaines comme un investissement rentable.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Avantages de la quantification. Les organisations, particulièrement celles qui disposent des ressources humaines et financières, ne devraient pas renoncer d'emblée au calcul de l'utilité, sous prétexte d'une application quelque peu complexe ; le recours au concept d'utilité, même imparfait, comporte plusieurs avantages. Premièrement, il fournit des **informations objectives** et irremplaçables permettant de prendre de meilleures décisions concernant l'introduction, le maintien ou le choix d'une méthode de sélection parmi plusieurs (Guion, 1998 ; Schneider et Schmitt, 1986). Dans un exemple tiré de Boudreau et Rynes (1885, cité dans Smith *et al.*, 1989), un directeur du personnel se demande laquelle des possibilités est la meilleure : 1) investir davantage dans une campagne de recrutement, 2) développer un instrument de sélection de type « données biographiques » ou 3) retenir les services d'une agence qui effectuerait un premier tamisage, puis conduire à l'interne une entrevue avec les candidats présélectionnés. Dans cette situation précise, le calcul de l'utilité a favorisé le recours à l'agence.

Signalons que le calcul de l'utilité n'est pas réservé aux seules méthodes de sélection du personnel. Il est déjà appliqué à d'**autres pratiques de gestion des ressources humaines** comme la formation, l'évaluation du rendement, les modes de rémunération et les formes d'implication des employés à la gestion (Boudreau, 1991 ; Huselid, Jackson et Schuler, 1997 ; Schmitt et Chan, 1998). Le calcul de l'utilité pourrait alors servir à évaluer les retombées de ces diverses pratiques afin d'investir dans les plus utiles pour la réalisation de la stratégie d'entreprise (Cascio, 1993). Bref, l'utilité est un cadre permettant aux gestionnaires de rassembler des informations nécessaires à la prise de décision et à leur justification.

Deuxièmement, le concept d'utilité permet de traduire la rentabilité d'une pratique de sélection sur le plan **monétaire**, apparemment si chère au monde organisationnel, concourant ainsi à rapprocher l'univers de la psychologie de celui des gestionnaires (Schuler et Guldin, 1991). L'idée de quantifier davantage la valeur des pratiques de gestion des ressources humaines n'est pas nouvelle (Gordon, 1972 ; Luthans et Maris, 1979 ; Fitz-Enz, 1980) ; plusieurs y voient un passage privilégié pour obtenir des ressources nécessaires ou une approbation auprès de la direction (Guion, 1998 ; Schneider et Schmitt, 1986).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Limites de la quantification. Mais attention, nous préviennent certains. Tout d'abord, la sophistication croissante des approches pour calculer l'utilité peut être **si complexe et si hermétique**, qu'elle en devient contre-productive (Schuler et Guldin, 1991). Nous l'avons vu, les gestionnaires sont souvent réticents, voire méfiants, à l'égard des approches qu'ils ne comprennent pas. Ensuite, plusieurs études ont mis à mal l'image du **gestionnaire rationnel**, qui prendrait ses décisions principalement à partir de données objectives, quantifiables et d'analyses coûts-bénéfices. En effet, Minztberg (1973, 1975) a démontré avec éloquence le caractère parfois intuitif des gestionnaires dont le processus décisionnel est souvent informel et subjectif (p. ex., approuver un projet sur la réputation de son promoteur plutôt qu'en fonction de ses qualités techniques). Les décisions concernant les pratiques de gestion des ressources humaines n'échappent sans doute pas à ce côté intuitif des gestionnaires.

À cet égard, une recherche de Latham et Whyte (1994) fournit une illustration particulièrement déconcertante du processus décisionnel des gestionnaires appelés à choisir une méthode de sélection. Éclairés par les conseils d'un psychologue, les gestionnaires étudiés ont été plus enclins à suivre la recommandation du psychologue lorsqu'elle était basée uniquement sur une étude de la validité, comparativement à une recommandation appuyée à la fois par une étude de la validité et un calcul de l'utilité. La propension des gestionnaires à se fier aux conseils du psychologue diminuait lorsque ce dernier ajoutait les résultats, pourtant objectifs et chiffrés, de l'analyse coûts-bénéfices provenant du calcul de l'utilité ! Les psychologues seraient-ils devenus plus rationnels que ne l'exige le monde organisationnel ? Quoique étonnant, ce résultat révèle que le gestionnaire demeure un être humain et que, pour le convaincre, il faut comprendre ce qui l'habite.

CONCLUSION

Le concept d'utilité, pour les chercheurs, représente une méthodologie rigoureuse permettant de chiffrer la rentabilité objective d'une intervention. Les résultats de leur démarche ont déjà marqué le domaine de la sélection du personnel et il ne saurait en être autrement pour les autres pratiques de gestion des ressources humaines.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

L'utilité, un mode de pensée. Mais pour les gestionnaires sur le terrain comme pour les professionnels qui interviennent dans les organisations, la principale contribution des approches visant à calculer l'utilité économique est l'élargissement du cadre d'analyse concernant les instruments de mesure. La valeur d'un instrument, de sélection ou autre, n'est pas que sa validité ; il y a d'autres aspects importants, d'autres critères. Et pour les identifier, il faut revenir aux raisons fondamentales pour lesquelles l'outil de mesure est utilisé.

L'utilité est avant tout pour le praticien un mode de pensée plus global, axé sur les objectifs visés et sur les résultats à atteindre, sur la différenciation entre les moyens et la fin, sur le fait, finalement, que rien n'est gratuit et que tout est relié. D'ailleurs, qu'est-ce que l'approche coûts-bénéfices, si ce n'est l'incarnation quantitative de ce mode de pensée ? Comment évaluer les coûts et les bénéfices sans obliger le décideur à réfléchir au problème à résoudre et à ses répercussions sur le fonctionnement de l'organisation, aux résultats à atteindre et aux moyens à prendre pour y arriver ? L'approche coûts-bénéfices, contrairement à ce que prétendent ses détracteurs, n'est pas nécessairement réductionniste. Bien au contraire, mais à condition de demeurer un mode de pensée, et non pas une finalité axée sur la quantification à tout prix. Car s'il y a risque de réduction, c'est dans le choix de paramètres plus faciles à quantifier au détriment de leur pertinence. Il faut parfois accepter des évaluations approximatives appliquées à de vrais enjeux plutôt que de mesurer avec précision un tableau incomplet.

Le gestionnaire des ressources humaines gagnerait à s'inspirer plus souvent de cette approche, ne serait-ce que pour avoir une vision plus axée sur les résultats et sur l'organisation, au lieu de n'être attentif qu'aux outils et aux techniques ; il ne doit pas renoncer à essayer de mesurer l'efficacité de ses interventions. Une telle quantification, selon Mercer (1989), est possible pour un bon nombre de pratiques touchant les ressources humaines ; il a écrit un livre sur la façon d'y arriver.

Il est vrai toutefois que l'utilité économique n'est pas tout dans une société. Il serait en effet navrant que la production efficace de biens et de services soit la seule finalité de la vie organisée. Des citoyens heureux et satisfaits dans leur travail représentent un objectif

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

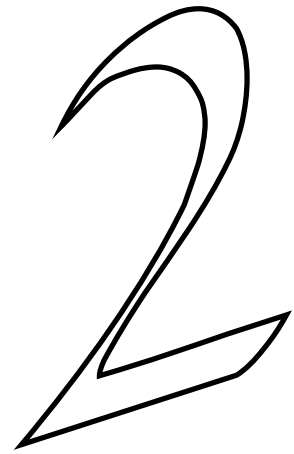
tout aussi, sinon plus, important. Aux dirigeants alors d'établir le difficile équilibre entre employés, clients et actionnaires. C'est la seule perspective valable à long terme.

Retombées concrètes de l'approche axée sur l'utilité. En plus du cadre général de pensée, ce chapitre a des apports très concrets. Premièrement, **quatre grands critères** pouvant servir à évaluer un instrument de sélection ont été dégagés : la validité bien sûr, mais aussi les coûts, la validité apparente et la facilité avec laquelle un instrument peut être défendu en cas de litige. Deuxièmement, la présentation des tables de Taylor-Russel a permis aux gestionnaires de connaître les **paramètres de la situation** qui influencent le potentiel d'un outil de sélection à améliorer le pourcentage d'employés satisfaisants après la sélection ; il s'agit de la validité, du taux de sélection et de la proportion de candidats susceptibles de convenir au poste. Les tables servent aussi à prédire l'accroissement du pourcentage d'employés satisfaisants découlant de l'application de l'outil de sélection. Troisièmement, grâce à l'approche de Brogden-Cronbach-Gleser et à ses diverses méthodes pour estimer la valeur monétaire des différences de rendement entre employés (S_y), nous savons qu'il est désormais possible d'évaluer l'amélioration produite par l'application d'un outil de sélection directement sur le plan des **gains économiques**. Cette approche n'est ni simple ni parfaite, mais elle est applicable lorsque nécessaire ; elle est grandement simplifiée par le recours à la méthode des « règles de proportion du salaire ou des biens et services produits ». Finalement, les résultats obtenus de ces études d'utilité économique sont de puissants instruments au service du spécialiste de la sélection. Il sait que ses outils sont **rentables** pour l'organisation, dans quelles conditions ils le sont davantage et il dispose de chiffres pour convaincre la direction, pour peu qu'elle soit rationnelle.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

C H A P I T R E



VALIDATION BASÉE SUR LE CONTENU DE L'INSTRUMENT DE MESURE

Un employé se voit refuser une promotion à un poste d'analyste-programmeur, parce qu'il a échoué les deux examens servant à évaluer ses compétences en informatique. Sa convention collective prévoit que *« le poste sera pourvu par le salarié qui a le plus d'ancienneté parmi ceux qui ont posé leur candidature, à la condition qu'il puisse satisfaire aux exigences normales de la tâche »*. Cet employé, le plus ancien à avoir postulé, soulève un grief et allègue qu'avoir échoué les examens ne prouve pas qu'il sera incapable de remplir le poste. Comment peut-on savoir si les résultats aux examens attestent réellement que l'employé ne satisfait pas aux exigences de l'emploi ?

Une municipalité désire mettre sur pied un processus de gestion et d'évaluation de la performance de son personnel cadre. Un des éléments clés du processus est de pouvoir évaluer leur performance de façon juste et équitable. Comment doit-elle s'y prendre pour choisir les critères qui serviront à l'évaluation de ses cadres ?

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Pour répondre à ces questions, il faut savoir jusqu'à quel point les résultats recueillis par ces instruments de mesure sont représentatifs du domaine à mesurer. Par exemple, les examens servant à évaluer les compétences en informatique pour le poste d'analyste-programmeur doivent porter sur les aspects importants de l'informatique réellement requis par le poste à pourvoir. Si le poste comporte principalement des tâches en analyse de système et en programmation, les examens doivent refléter ces exigences, avec le même degré de difficulté que celui rencontré dans les tâches. Il en est de même pour les critères servant à évaluer la performance du personnel cadre : ils doivent couvrir l'ensemble des aspects importants de leur travail, qu'il s'agisse de tâches, de résultats ou de compétences. La validation basée sur le contenu est une démarche qui vise précisément à évaluer la pertinence du contenu d'un instrument de mesure pour un usage donné.

La validation basée sur le contenu est une notion fondamentale pour trois raisons. En gestion des ressources humaines, la validation basée sur le contenu est devenue un enjeu considérable. En effet, depuis une vingtaine d'années, on assiste à une recrudescence de l'intérêt pour cette forme pourtant ancienne de validation, autant de la part des spécialistes de la mesure que des gestionnaires de personnel (voir Goldstein, Zedeck et Schneider, 1993). C'est une notion incontournable, même pour ceux qui en ignorent le nom.

Trois raisons semblent expliquer cet état de fait. Premièrement, il y a une augmentation croissante des **procédures judiciaires** en matière de gestion des ressources humaines ; le Canada n'échappe pas à ce mouvement, amorcé principalement aux États-Unis dans les années 1960. Les organisations prennent de plus en plus conscience de l'importance de pouvoir démontrer hors de tout doute raisonnable que les instruments sur lesquels ils fondent leurs décisions sont pertinents à l'emploi et ne portent aucune atteinte aux droits des personnes, que ce soit en vertu des chartes, des lois du travail ou de toute autre politique gouvernementale. Or, devant les tribunaux et les organismes qui édictent les lignes directrices en matière d'instruments de mesure, la validation basée sur le contenu est considérée maintenant comme une démarche de validation en soi, acceptable et suffisante, et non plus simplement comme une solution de remplacement inférieure aux autres formes de validation (Arvey et Faley, 1988). Qui plus est, selon des jugements rendus par divers tribunaux,

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

spécialement en contexte de sélection et de promotion du personnel, il est clair que la validation basée sur le contenu est l'un des aspects qui retiennent particulièrement l'attention des magistrats ; il sera beaucoup plus ardu de défendre un outil de mesure si cette démonstration est faible (Kleiman et Faley, 1985 ; Landy et Vasey, 1991). Les utilisateurs et les candidats, aussi, acceptent plus facilement un instrument fondé sur la validation basée sur le contenu (Arvey et Faley, 1988) que sur des démarches statistiques plus abstraites à leurs yeux. Une décision récente du Tribunal des droits de la personne du Québec (1999) est de nature à inciter les spécialistes des ressources humaines sous sa juridiction à prendre au sérieux l'importance de pouvoir réaliser une telle démarche de validation[1].

Deuxièmement, au cours des années 1970, on a pu observer une baisse de confiance envers les tests psychométriques comme outils de sélection (p. ex., tests d'aptitudes mentales ou inventaires de personnalité). Il faut dire que les résultats de recherche concernant la validité des tests psychométriques étaient décevants et que ces tests étaient, de surcroît, fort vulnérables sur le plan légal. Il se peut aussi que des mouvements idéologiques et sociaux aient contribué à leur déclin momentané. Peu importe les raisons, d'autres instruments de sélection et d'évaluation, plus proches du contenu réel des tâches à exécuter en emploi, ont pris la relève. Les analyses de cas, les jeux de rôle, les simulations de travail en équipe, les épreuves du courrier (*In-Basket Tests*) et autres mises en situation occupent maintenant une grande place dans les services de ressources humaines et dans les cabinets spécialisés. Même les entrevues ont été transformées pour être centrées davantage sur les postes à pourvoir. L'usage accru de ces instruments de sélection, centrés sur l'emploi, participe au retour en force de la validation basée sur le contenu, parce qu'ils sont presque tous développés à partir de cette approche. En effet, la validation basée sur le contenu est quasiment indissociable de l'**élaboration d'instruments de mesure**, surtout en ce qui a trait aux outils mesurant des attributs concrets et observables comme les connaissances et les habiletés relatives au monde du travail.

1. Dossier 505-53-000008-976, Commission des droits de la personne et des droits de la jeunesse contre Institut Demers inc.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Troisièmement, l'élaboration des instruments de mesure, que ce soit en sélection ou en promotion du personnel, en évaluation du rendement, en formation ou autre, est du ressort des professionnels en ressources humaines ; la validation basée sur le contenu offre un outil indispensable pour s'acquitter de cette responsabilité. C'est la seule approche en matière d'élaboration d'instruments de mesure qui soit à la fois rigoureuse, aisément défendable en cas de litige et **accessible** aux professionnels en ressources humaines, sans exiger pour autant qu'ils soient spécialistes en mesure et évaluation ou férus de statistiques. C'est une approche fondée sur la logique du gros bon sens, comportant peu de complexités techniques ou statistiques. Un tel processus d'élaboration d'instruments est présenté plus loin dans cet ouvrage. Pour l'instant, voyons en quoi consiste la notion de validation basée sur le contenu.

Plan du chapitre. Ce chapitre comporte cinq sections. La première introduit le concept de validité pris au sens large, tout en soulignant l'implication de cette notion pour la gestion des ressources humaines. Les sections suivantes sont consacrées à la validation basée sur le contenu. La deuxième présente les notions de base et leurs définitions en relation avec le monde du travail. La troisième section donne les différentes composantes d'un instrument de mesure et de la prise de décision, car chacune d'elles doit être intégrée à un processus de validation basée sur le contenu. La quatrième section traite du principe de la représentativité des composantes d'un instrument de mesure et fait état de ce qu'il advient si des composantes ne reflètent pas le domaine de l'emploi. Enfin, la cinquième partie énumère les conditions propices à l'application d'une telle démarche de validation.

VALIDITÉ, VALIDATION ET GESTION DES RESSOURCES HUMAINES

Avant d'examiner de plus près la validation basée sur le contenu, il est indispensable de clarifier le concept général de validité et d'en saisir les différentes caractéristiques. Les lignes qui vont suivre ne sont pas de nature à capter l'attention d'un gestionnaire qui se réclame le moindrement pragmatique. Il n'est pas facile pour qui n'est pas familier avec ces concepts d'en percevoir au premier abord toute la

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

portée pratique. Néanmoins, quoi que l'on puisse en comprendre ne change en rien la réalité : le concept de validité est toujours en cause lorsque des personnes, des objets ou des situations sont évalués, appréciés, mesurés. Et de telles mesures sont constamment requises dans la gestion contemporaine des ressources humaines et ses nombreuses ramifications avec le droit du travail : sélection, promotion, gestion du rendement, mesures disciplinaires, diagnostic des besoins de formation et suivi, presque toutes les formes de rémunération au mérite, etc.

VALIDITÉ

La qualité primordiale d'une mesure est sa validité ; c'est le principal aspect à considérer lorsque vient le temps d'évaluer un test ou tout autre instrument de mesure (Guion, 1998). La validité est la capacité d'un instrument à mesurer ce qu'il est censé mesurer (Nunnally et Bernstein, 1994) ou, d'un point de vue plus général, à atteindre les objectifs qui lui sont assignés (Carmines et Zeller, 1979 ; McCormick et Tiffin, 1974). Ainsi, un examen de français est valide s'il mesure vraiment la compétence en français, et seulement cette compétence ; une entrevue de sélection est valide si elle permet de prédire la performance des candidats, une fois embauchés ; un système d'évaluation du rendement est valide dans la mesure où la contribution de chaque employé est cotée équitablement et sans biais, en fonction des exigences normales de son poste.

Bien que l'on entende souvent dire que tel test ou tel instrument est valide, cette formulation est incorrecte. À strictement parler, la validité ne porte pas sur l'instrument de mesure en lui-même, mais plutôt sur les résultats obtenus et leurs usages. En effet, *la validité évalue jusqu'à quel point les* ***inférences****, les interprétations faites à partir des résultats recueillis à un instrument de mesure sont exactes et fondées dans la situation particulière où elles sont appliquées* (Nunnally et Bernstein, 1994 ; Schmitt et Chan, 1998 ; *Standards*, 1999). Par exemple, lorsqu'une entrevue de sélection est utilisée pour choisir les meilleurs candidats, on infère (on interprète) que le résultat à cette entrevue indique si le candidat pourra ou non répondre de façon satisfaisante aux exigences du poste. Lorsqu'un chef d'équipe évalue le rendement d'un de ses membres, on fait l'inférence que les résultats attribués reflètent exactement la performance de ce membre. Bref, la validité

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

porte sur les résultats obtenus à l'instrument ou, plus précisément, sur l'interprétation de ces résultats en fonction d'un usage particulier (Binning et Barrett, 1989 ; Cronbach, 1971 ; Guion, 1998).

DEUX SORTES D'USAGES, DES INFÉRENCES (INTERPRÉTATIONS) DESCRIPTIVES OU RELATIONNELLES

Les usages que l'on peut faire des résultats (scores) obtenus à un instrument de mesure se classent en deux groupes, selon les inférences formulées : 1) des inférences descriptives, à savoir inférer *ce qui* est mesuré par les résultats obtenus, ou 2) des inférences relationnelles, à savoir inférer *avec quoi* ces résultats sont reliés.

Inférences descriptives. Le premier type d'inférences, que Guion (1998) appelle descriptives ou interprétatives, porte sur les résultats comme mesures valables de l'objet ou de la caractéristique censé être mesuré par l'instrument. Par exemple, si un examen de conduite est utilisé pour sélectionner des camionneurs, on fait l'inférence qu'un résultat élevé à l'examen montre que le candidat maîtrise les aspects évalués de la conduite d'un véhicule lourd. Pour que l'inférence soit exacte, il ne faut pas que les résultats soient faussés par des facteurs externes, comme la nervosité du candidat en présence de l'examinateur ou sa difficulté à comprendre les directives. Si une entrevue est employée pour évaluer chez un candidat son habileté à résoudre des problèmes de gestion, on présume que l'évaluation du comité de sélection est conforme au niveau d'habileté du candidat. Si un inventaire de personnalité, utilisé lors d'un processus de sélection, est destiné à mesurer la conscience professionnelle, on infère que le résultat obtenu traduit réellement le comportement du candidat, et non pas son désir de bien paraître aux yeux de l'examinateur pour obtenir le poste. Si l'on procède à l'évaluation annuelle du rendement d'un employé, on fait l'inférence que les résultats obtenus reflètent sa performance, en conformité avec sa description de tâches, les objectifs de son service et tout autre engagement particulier. L'évaluation doit être exempte des biais de l'évaluateur et des falsifications de la part de l'évalué.

Ce premier type d'inférences a trait à la nature **intrinsèque** de la mesure. En évaluer la validité, c'est essentiellement chercher à savoir *jusqu'à quel point les évaluations ou les résultats obtenus mesurent bien la caractéristique ou l'objet visé par l'instrument, et rien d'autre*. Si

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

ces inférences sont valides, elles permettent de décrire avec assurance l'objet mesuré par l'instrument et ainsi d'interpréter raisonnablement les résultats obtenus en fonction de cet objet, et seulement lui. Par exemple, les résultats à cet examen de conduite indiquent le degré de compétence par rapport aux aspects mesurés ; ou bien, les évaluations du rendement sont conformes à ce que l'employé a réalisé durant la période évaluée. Étant donné que la validité des inférences descriptives dépend en grande partie du processus suivi lors de l'élaboration de l'instrument de mesure, Guion (1998) propose de l'appeler la **validité psychométrique** (p. 18, 238).

Inférences relationnelles. Le deuxième type d'inférences que l'on peut faire lorsqu'on a recours à un instrument de mesure consiste à vouloir connaître les résultats d'une autre variable, qui n'est pas mesurée directement par l'instrument, mais qui lui est reliée. Reprenons l'exemple de l'examen de conduite utilisé pour sélectionner des camionneurs. En plus de faire une inférence descriptive concernant le degré de compétence attesté par les résultats obtenus, on peut aussi faire une inférence relationnelle par rapport au bilan routier qu'auront les candidats une fois en poste. On présume par exemple qu'un résultat élevé à l'examen se traduira éventuellement par un faible taux d'accidents ou de contraventions, par des coûts d'entretien mécanique moins élevés, ou par tout autre indicateur d'efficacité. Lorsqu'un test d'aptitudes mentales est utilisé pour évaluer la capacité de compréhension et d'apprentissage, l'organisation est surtout intéressée par le fait que les résultats obtenus permettent de repérer lesquels parmi les candidats seront les plus efficaces dans leurs tâches. Lorsqu'un interviewer évalue la motivation d'un candidat en se basant sur ses comportements durant une entrevue (poignée de main dynamique, réponse rapide, posture droite, etc.), il fait l'inférence (la plupart du temps à tort…) que de tels comportements sont un gage d'efforts soutenus et de persévérance dans l'emploi.

Ce type d'inférences, qualifiées de relationnelles (Guion, 1998), concerne la nature **extrinsèque** de la mesure ; ici, l'instrument est **prédicteur d'une autre variable**. La validité de ces inférences traduit *jusqu'à quel point les évaluations ou les résultats obtenus sont reliés à une autre variable (caractéristique personnelle, compétence ou rendement) extérieure à l'instrument lui-même*. Si elles sont valides, ces inférences autorisent l'interprétation des résultats par rapport aux résultats correspondants au sujet d'une autre variable, différente mais reliée. Dans

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

le cas du candidat camionneur, plus les résultats de son examen de conduite sont élevés, meilleur sera son bilan routier. Plus ce candidat a démontré de bonnes aptitudes mentales au test, plus il sera efficace dans son emploi.

LES STRATÉGIES TRADITIONNELLES DE VALIDATION

La validité des inférences, descriptives et relationnelles, peut être évaluée de diverses manières. Nous allons maintenant présenter sommairement les grandes approches pour démontrer cette validité ; deux d'entre elles seront vues en détail par la suite, dans ce chapitre et dans le suivant.

Validation. La validation est le processus par lequel sont accumulés les évidences ou les éléments de preuve relativement à la validité des inférences (*Standards*, 1999). Vouloir connaître la validité des résultats à un instrument de mesure dans une situation donnée, c'est établir en quelque sorte un processus de contrôle, processus requis dans toute gestion rationnelle orientée vers la réalisation d'objectifs clairs. Depuis plusieurs décennies, les principales méthodes de validation ont été regroupées en trois grandes stratégies : la validation basée sur le contenu de l'instrument de mesure, la validation basée sur la relation avec d'autres variables et la validation basée sur le construit mesuré par l'instrument. Elles sont aussi appelées, plus familièrement, validation de contenu, validation critériée et validation de construit.

Validation basée sur le contenu. La stratégie de validation basée sur le contenu cherche à démontrer jusqu'à quel point les composantes d'un instrument de mesure (items, questions, tâches, directives, processus de correction, etc.) permettent de recueillir des résultats qui soient représentatifs du domaine de contenu visé, à savoir l'objet ou la caractéristique à mesurer. Par exemple, est-ce que les questions d'un examen d'anglais écrit mesurent les compétences en anglais (le domaine de contenu) préalablement définies et seulement elles ? Cette stratégie donne lieu à une évidence de validité basée sur la **représentativité** de ce qui est mesuré par l'instrument (Kerlinger, 1986).

Validation basée sur la relation avec d'autres variables. La stratégie de validation basée sur la relation avec d'autres variables recherche les éléments de preuve qui confirment que les résultats à

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

un instrument sont systématiquement reliés à une variable externe à cet instrument. Les résultats à l'instrument sont utilisés, non pas pour eux-mêmes, mais en tant qu'indicateurs de cette autre variable, appelée critère, qui est le véritable intérêt de l'utilisation de l'instrument. Par exemple, est-ce que les résultats obtenus à un test d'aptitudes mentales par des candidats lors du processus de sélection sont reliés à leur rendement au travail (le critère) dans leur nouvel emploi ? La validité est examinée sous l'angle de la **relation** entre les résultats à l'instrument de mesure et une autre variable.

Validation basée sur le construit mesuré. Un construit est une caractéristique psychologique ou un objet abstrait qui ne peut pas être observé directement, mais qui se manifeste à travers divers comportements. Le construit est en quelque sorte une hypothèse que les chercheurs « construisent » pour expliquer des comportements qui sont reliés entre eux (Nunnally et Bernstein, 1994). Par exemple, lorsqu'une personne se sent constamment fatiguée au travail, qu'elle n'a plus confiance en elle, qu'elle est devenue très sensible à la critique, qu'elle a de la difficulté à prendre des décisions, qu'elle est de moins en moins efficace, qu'elle en perd le sommeil et l'appétit, on dit de cette personne qu'elle souffre d'épuisement professionnel (*burnout*). La stratégie de validation basée sur le construit tente de démontrer si les résultats à un instrument de mesure reflètent réellement le construit devant être mesuré. Par exemple, est-ce que ce questionnaire portant sur l'épuisement professionnel mesure bien et sans biais cet état psychologique ? L'évidence de validité ainsi obtenue repose sur **l'explication** des résultats des répondants à l'instrument de mesure par la caractéristique à mesurer, et seulement elle (Kerlinger, 1986).

Dans la dernière version du document *Standards for Educational and Psychological Testing* (1999), produit conjointement par l'American Educational Research Association, l'American Psychological Association et le National Council on Measurement in Education, on propose une nouvelle classification des méthodes de validation. D'abord, on ne parle plus de validation basée sur le construit parce que, dans les faits, cette approche comprend les autres méthodes de validation. On élimine ainsi une certaine redondance, qui était parfois source de confusion relativement à cette forme de validation. En revanche, on ajoute deux autres stratégies, soit la validation basée sur la structure interne de l'instrument et la validation basée sur le

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

processus suivi par les répondants pour répondre à l'instrument. Comme ces dernières formes de validation ne cadrent pas avec l'objectif du présent ouvrage, ils ne feront pas l'objet d'une présentation[2].

Types d'inférences et stratégies de validation. Il existe un lien évident entre les stratégies de validation et le type d'inférences à valider. Ainsi, la stratégie de validation basée sur le contenu est particulièrement appropriée aux inférences **descriptives**, dont la validité repose sur le fait que les résultats mesurent vraiment la caractéristique ou l'objet visé par l'instrument. En attendant des explications plus poussées, on peut dire que la validation basée sur le contenu est surtout utile lorsque les instruments mesurent des comportements, des connaissances ou des habiletés directement observables. Si les instruments mesurent des caractéristiques plutôt abstraites (la capacité d'apprentissage, la motivation, le leadership, etc.), il faudra alors recourir à d'autres stratégies, notamment à la validation basée sur la structure interne de l'instrument ou à celle basée sur le processus de réponse. Lorsqu'il s'agit d'évaluer la validité d'inférences **relationnelles**, c'est-à-dire jusqu'à quel point les résultats obtenus sont reliés à une variable extérieure à l'instrument de mesure, la validation basée sur la relation à d'autres variables est alors la stratégie qu'il convient d'adopter.

PRÉCISIONS ADDITIONNELLES

La validité est un concept unitaire. Traditionnellement, la validité était qualifiée du nom de la démarche de validation employée. Ainsi, la validation basée sur le contenu donnait lieu à une évidence de **validité de contenu,** la validation basée sur la relation à d'autres variables à une évidence de **validité critériée** et la validation basée sur le construit à une évidence de **validité de construit**[3]. Cette

2. On propose également dans cet ouvrage de désigner par le terme « construit » tout concept ou caractéristique mesuré par un test, que ce concept ou cette caractéristique soit observable directement ou non (*Standards*, 1999, p. 5).
3. Il existe d'autres formes de validité, la plupart assimilables à la validation basée sur la relation avec d'autres variables ; les plus connues sont la validité synthétique, convergente, discriminante et incrémentielle. Le concept de validité apparente, dont on entend souvent parler et qui a son importance en matière d'instruments de mesure n'est pas, à proprement parler, un aspect de la validité. La validité incrémentielle est abordée plus loin.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

pratique tend à disparaître, du moins chez les universitaires, afin de mettre en évidence le caractère entier de la validité. En effet, bien qu'il existe diverses stratégies de validation qui éclairent des aspects différents de la validité, ceux-ci ne représentent pas des types de validité distincts. La validité est un concept unitaire ; elle concerne l'exactitude des inférences faites à partir des résultats à un instrument de mesure dans une situation donnée. La validité s'intéresse à l'accumulation de tous les éléments de preuve, peu importe la démarche de validation utilisée, qui contribuent à établir le bien-fondé de ces inférences (*Standards*, 1999). Les diverses stratégies de validation se complètent (Nunnally et Bernstein, 1994), chacune mettant en évidence un élément de cette preuve ; ainsi le processus idéal de validation devrait intégrer tous les éléments de preuve nécessaires à l'usage prévu des résultats (*Standards*, 1999)[4].

4. En fait, on se rend de plus en plus compte que le processus de validation et le processus général de développement des théories convergent vers les mêmes logiques de base de la recherche empirique : définition des variables (ou des construits), élaboration d'instruments pour les mesurer et étude des relations entre ces variables (voir Schmitt et Landy, 1993). Binning et Barrett (1989) démontrent la relation étroite entre les stratégies de validation et le processus de développement des théories. Une telle intégration élargit la compréhension du processus de validation. Comme pour toute question de recherche et d'acquisition des connaissances, il ne faut pas se laisser limiter par des conventions étroites ; il n'y a pas de méthodologie rigide et idéale en matière de validation (Schuler et Guldin, 1991). Les méthodologies utilisées peuvent être innombrables, comme cela est admis en recherche. De la même façon, il est rare qu'une seule étude, si bien menée soit-elle, permette de tirer des conclusions définitives lorsque la question de recherche est le moindrement importante et complexe. Guion (1998) va plus loin en préférant aborder la question sous l'angle beaucoup plus large de l'évaluation des tests et de leurs usages, sans nécessairement faire référence au concept de validité. Voilà une manière originale de ne pas tomber dans les ornières de trois validités distinctes et séparées. La dernière version des *Standards* (1999) s'inscrit résolument dans cette vision intégrative où tous les éléments de preuves doivent former un tout cohérent participant à la démonstration de la validité des inférences issues des résultats. La présentation qui en est faite pour le contexte de la gestion des ressources humaines met en relief la complémentarité entre la validation basée sur le contenu et celle basée sur la relation avec d'autres variables (p. 153-155).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Le caractère unitaire de la validité est incontestable. Par contre, il est parfois plus commode de faire directement référence tantôt à la validité de contenu, tantôt à la validité critériée. Il faudra alors comprendre qu'il s'agit de la validité examinée sous l'angle du contenu de l'instrument ou sous celui de la relation avec une autre variable.

La validité est une question de degré. Il est incorrect de dire qu'un instrument est valide. D'abord, on se rappelle que la validité concerne les résultats recueillis par l'instrument ou, plus précisément, les inférences qu'on peut en tirer. Ensuite, la validité n'est pas une caractéristique dont l'évaluation se divise en deux : valide ou invalide. Elle varie plutôt suivant un continuum, qui va de l'absence totale de validité à une validité parfaite, en passant par tous les niveaux intermédiaires. Par conséquent, lorsqu'on désire exprimer la validité des résultats obtenus à un instrument de mesure, il vaut mieux qualifier le degré de cette validité, idéalement par un indice statistique (dans la plupart des cas, le coefficient de corrélation). Par exemple, la validité des résultats est de 0,35 dans telle situation (corrélation obtenue par une stratégie de validation basée sur la relation avec une autre variable). Signalons que la validité parfaite n'existe pas, du moins pas pour les instruments utilisés en gestion des ressources humaines. Par ailleurs, il est rare que les résultats obtenus à un instrument soient totalement invalides.

La validité s'applique à chaque usage des résultats dans une situation donnée. Un instrument peut avoir un certain degré de validité pour un usage et ne pas en avoir pour un autre (Nunnally et Bernstein, 1994). Par exemple, les résultats à un examen d'informatique peuvent être modérément valides pour la sélection de techniciens dans une entreprise et être totalement invalides pour la sélection de concepteurs de systèmes dans une autre entreprise. Il est possible de généraliser la validité d'une situation à une autre, mais seulement lorsque des données ou des études le permettent (voir chapitre 3, à la section « Méta-analyse, généralisation de la validité et autres méthodes de validation »).

RESPONSABILITÉ DU PROFESSIONNEL EN RESSOURCES HUMAINES

Importance de recourir à des instruments valides. Le concept de validité constitue la pierre angulaire de toute activité scientifique ou professionnelle concernant la mesure et l'évaluation (Sarrazin, 1994).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

L'efficacité des instruments de mesure dépend fondamentalement de la validité des inférences faites à partir des résultats obtenus au moyen de ces instruments. Si les inférences ne sont pas valides, les résultats ne peuvent pas servir à décrire le répondant (inférence descriptive) ni à prédire son comportement (inférence relationnelle). Un instrument dont la validité est insuffisante n'a pas sa place en gestion des ressources humaines. Non seulement il ne peut servir à améliorer les décisions en matière de gestion, mais en outre il peut être nuisible et entraîner des conséquences néfastes : choix de candidats inadéquats lors d'embauche ou de promotion, évaluation inéquitable des employés, difficulté à se défendre en cas de grief ou de poursuite devant les tribunaux, sans parler du manque d'éthique à l'égard des personnes injustement traitées.

Responsabilité du concepteur et de l'usager. Nous devons comprendre quelle est la responsabilité de l'usager et du concepteur au regard de la validité des instruments. Prenons l'exemple d'un agent de dotation qui, pour la sélection de vendeurs, décide d'employer un inventaire de personnalité censé mesurer la confiance en soi. Celui qui a conçu cet inventaire doit établir la preuve de ses prétentions en fonction des usages qu'il propose. À cet égard, la validité des inférences descriptives (validité psychométrique) est d'abord la responsabilité du concepteur de l'instrument de mesure ou de son fournisseur. Quant à l'utilisateur, il a le devoir de justifier la pertinence de l'instrument par rapport à l'usage particulier qu'il en fait. Ainsi, l'agent de dotation doit pouvoir démontrer que la confiance en soi est effectivement une qualité importante chez les vendeurs concernés et que l'inventaire qu'il a choisi pour la mesurer convient à la situation. Il est clair que la preuve de la validité des inférences descriptives et relationnelles propres à sa situation lui incombe. Lorsque le professionnel chargé des ressources humaines est à la fois le concepteur et l'utilisateur d'un instrument (entrevue structurée, examen de connaissances, simulation, etc.), il a l'entière responsabilité de la validité des résultats (Guion, 1998 ; *Standards*, 1999, p. 11 et les articles 6.3, 6.4, 6.5, 6.15, 11.1, 11.2 et 11.4).

Stratégies de validation les plus courantes en gestion des ressources humaines. Il est essentiel qu'un gestionnaire en ressources humaines, et surtout un spécialiste en sélection et en évaluation du personnel, soit familier avec le concept de validité ; il doit savoir comment assurer ou vérifier la validité d'un instrument de mesure.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Deux stratégies de validation sont importantes pour le gestionnaire de ressources humaines : la validation basée sur le contenu de l'instrument de mesure, qui fait l'objet des prochaines sections, et la validation basée sur la relation avec d'autres variables, qui est abordée au chapitre suivant.

VALIDATION BASÉE SUR LE CONTENU ET APPLIQUÉE AU MONDE DU TRAVAIL

La validité d'un instrument de mesure repose sur la représentativité de son contenu par rapport à ce qui est supposé être mesuré par cet instrument. Par exemple, un examen de fin de session devrait porter sur la matière vue au cours, et seulement sur cette matière. Comme cet examen est conçu pour mesurer directement la performance scolaire dans ce cours et qu'il n'est pas utilisé pour évaluer quelque chose d'autre, l'examen doit se défendre par lui-même et les résultats doivent refléter directement ce que l'examen est censé mesurer. Il s'agit d'un usage mettant en cause une inférence descriptive dont l'exactitude relève de la validation basée sur le contenu. Cette démarche est plus simple en théorie qu'en pratique, surtout en gestion des ressources humaines où son application devient vite complexe. C'est ce que nous verrons dans la suite de ce chapitre.

DÉFINITION GÉNÉRALE

Représentativité par rapport au domaine à mesurer. Un instrument de mesure, que ce soit un test, un examen, une entrevue ou une simulation, ne sert qu'à recueillir un échantillon de réponses ou à observer un échantillon de comportements appartenant à un domaine préalablement défini (Guion, 1965, 1977). La validation basée sur le contenu examine la pertinence du contenu, de la matière, des thèmes d'un instrument de mesure par rapport à ce domaine (Kerlinger, 1986 ; *Standards*, 1999). Elle évalue *jusqu'à quel point le contenu de l'instrument de mesure (items, questions, tâches, procédures d'administration et de correction, etc.) permet de recueillir des résultats représentatifs du domaine à mesurer, seulement lui, et si ce domaine est pertinent à l'usage prévu des résultats*. Par exemple, on s'attend à ce qu'un examen de français écrit, pour faire l'embauche de personnel de bureau, soit composé de questions portant sur les aspects importants

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

de l'orthographe, de la syntaxe, de la grammaire et peut-être aussi de la stylistique. Toutefois, on serait surpris d'y retrouver des questions sur des règles de mise en page ou de rédaction des références bibliographiques. On compare le contenu de l'examen non pas à ce que l'on croit être le domaine du français écrit en général, mais au domaine du français écrit requis par l'emploi considéré. Ainsi, le contenu de l'examen peut être valide par rapport au domaine du français écrit, mais ne pas l'être dans ce contexte précis de sélection du personnel, si les compétences mesurées en français ne sont pas nécessaires à ces emplois de bureau.

Basé sur le jugement. La validation basée sur le contenu est un processus généralement fondé sur le jugement. Les méthodes utilisées font intervenir des experts qui évaluent la concordance entre les composantes de l'instrument de mesure et les éléments du domaine à cerner. Ces experts sont le plus souvent des spécialistes reconnus du domaine mesuré ; en anglais, on les désigne par l'expression « *subject matter experts* » (SME). Par exemple, des experts en comptabilité peuvent être amenés à se prononcer sur la représentativité de l'ensemble des problèmes et de leur solutionnaire faisant partie des examens pour devenir membre de l'Ordre des comptables agréés. Ils devront démontrer, par raisonnement logique plutôt que par des statistiques, que l'instrument de mesure représente raisonnablement le domaine visé.

Les jugements ne sont pas absolus (Murphy et Davidshofer, 1988) et les experts ne s'accordent pas toujours dans leur appréciation de la pertinence du contenu d'un instrument de mesure. Néanmoins, leurs jugements ne sont pas arbitraires. La validité de l'instrument doit être évaluée en comparant systématiquement le contenu de l'instrument avec celui du domaine à mesurer, préalablement défini avec soin.

Le domaine de contenu doit être spécifié. La validation basée sur le contenu exige donc que l'on puisse spécifier le domaine de contenu que l'instrument est censé représenter, c'est-à-dire l'ensemble de tous les éléments pertinents en rapport avec ce domaine (Kerlinger, 1986 ; Nunnally et Bernstein, 1994). Pour évaluer la validité des examens de l'Ordre des comptables agréés, les experts consultés devront d'abord s'entendre sur le domaine de la comptabilité à

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

évaluer : les champs de pratiques privilégiés par l'Ordre, les connaissances nécessaires à ces pratiques, le niveau de qualité attendu, les différents contextes d'utilisation possibles, etc.

Voyons un autre exemple où une organisation en commerce de détail désire assurer la validité des critères servant à évaluer le rendement des caissiers et caissières. Il faudra d'abord identifier les aspects importants de cet emploi, à l'aide de personnes compétentes en la matière (*subject matter experts*), notamment des caissiers d'expérience, des superviseurs ou même des clients. Les aspects identifiés constituent le domaine de contenu. Il pourrait contenir des dimensions comme 1) l'usage de la caisse enregistreuse (vitesse et précision), 2) la relation avec la clientèle, 3) le travail en équipe ou 4) le respect des directives et politiques de l'organisation. Ensuite, il faudra vérifier si les critères d'évaluation représentent exactement ces aspects de l'emploi, et pas autre chose.

DÉTERMINATION DU DOMAINE DE CONTENU EN CONTEXTE DE GESTION DES RESSOURCES HUMAINES

En gestion des ressources humaines, le domaine de contenu s'établit par rapport à l'emploi, au regard des tâches, des comportements, des connaissances ou des habiletés. Il s'agit alors de vérifier si le contenu de l'instrument constitue un échantillon représentatif de ces tâches, de ces comportements, de ces connaissances ou de ces habiletés. Plus les items et les tâches de l'instrument de mesure correspondent au contenu de l'emploi, plus la preuve de validité basée sur le contenu est convaincante. Par exemple, il serait aisé d'établir la validité d'un examen de soudure au cours duquel les candidats seraient invités à effectuer le même type de soudure, avec le même équipement, dans les mêmes conditions et avec les mêmes exigences que dans la situation réelle de travail. Il en serait de même d'une entrevue de sélection où des problèmes réels déjà survenus dans ce poste de travail seraient soumis aux candidats.

Spécifié par les comportements dans l'emploi, par les résultats produits ou par les connaissances ou les habiletés nécessaires à ces comportements. La United States Civil Service Commission recommande de limiter le domaine de contenu aux seuls aspects reliés à la **performance** dans l'emploi envisagé (Gavin, 1977, p. 11). C'est

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

également le point de vue de Hunter (1981). La Society for Industrial and Organizational Psychology (SIOP) précise que le domaine de contenu d'un outil de sélection et de promotion du personnel doit être 1) la **performance** (rendement) dans l'emploi pour lequel les candidats sont évalués, ou 2) **les connaissances ou les habiletés nécessaires à cette performance** (1987, p. 19). Selon *The Uniform Guidelines on Employee Selection Procedures*, les résultats des candidats à un instrument de sélection doivent constituer un échantillon représentatif 1) des **comportements** dans l'emploi en question (*behaviors*), 2) des **résultats produits** par ces comportements (*work product*) ou 3) des **connaissances** ou **habiletés** requises pour l'exécution de ces comportements (Equal Employment Opportunity Commission *et al.*, 1978, section N° 1607.14-C-4). Dans ce dernier cas, l'expression « performance » a été remplacée par les « comportements » et les « résultats produits par ces comportements ». Finalement, les *Standards* mentionnent que le domaine de contenu doit porter sur les **tâches** ou sur les **connaissances**, les **habiletés**, les **aptitudes** ou les **autres caractéristiques** personnelles (1999, article 14.8).

On en conclut que, *en gestion des ressources humaines, la validité d'un instrument doit être établie par rapport aux comportements, aux résultats produits par ces comportements, ou aux connaissances ou habiletés nécessaires à l'exécution de ces comportements, et seulement par rapport à ces éléments*.

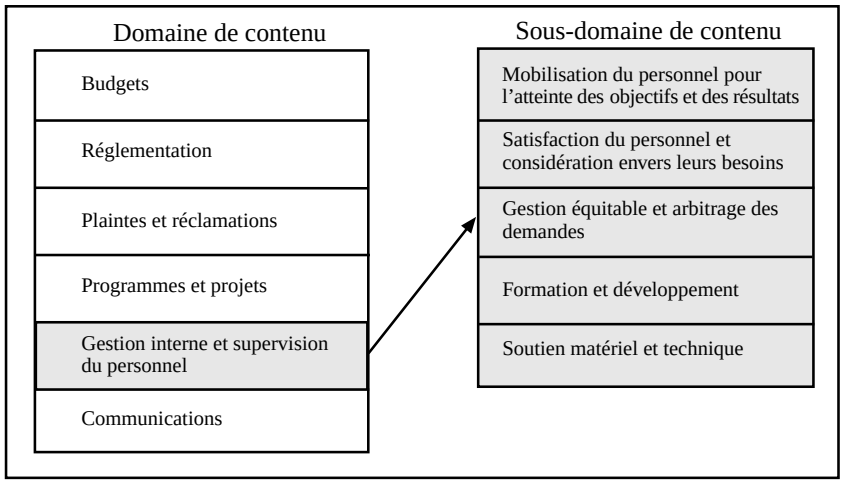
DOMAINE ET SOUS-DOMAINE DE CONTENU

En guise de spécification du domaine de contenu pour un emploi de directeur municipal, les fonctions suivantes ont été retenues : 1) gestion des budgets, 2) gestion de la réglementation, 3) gestion des plaintes et des réclamations, 4) gestion des programmes et des projets, 5) gestion interne et supervision du personnel et 6) gestion des communications (voir figure 2.1). Chacune de ces fonctions peut être considérée à son tour comme un sous-domaine de contenu, puis être définie de manière plus détaillée. Par exemple, la fonction de gestion interne et de supervision du personnel peut devenir un sous-domaine défini par les aspects suivants : 1) mobilisation du personnel pour l'atteinte des objectifs et des résultats, 2) satisfaction du personnel et considération envers leurs besoins, 3) gestion équitable et arbitrage des demandes, 4) formation et développement, 5) soutien matériel et

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 2.1
EXEMPLE DE DOMAINE DE CONTENU ET D'UN SOUS-DOMAINE POUR UN EMPLOI DE DIRECTEUR MUNICIPAL



technique. Un instrument de mesure peut être conçu pour évaluer le domaine de contenu en entier ou en partie. Gavin (1977) rappelle, à cet égard, la distinction entre la validité d'un instrument de mesure, qui peut ne couvrir qu'un nombre restreint des sous-domaines de l'emploi, et la validité de l'ensemble du processus de sélection, qui devrait représenter tout le domaine de contenu de l'emploi[5].

5. Certains auteurs font une distinction entre les concepts de domaine et d'univers. Selon Gavin (1977), l'univers de contenu est l'ensemble exhaustif de tous les éléments constituant le contenu visé. Dans le cadre d'un emploi, l'univers de contenu de l'emploi (*job content universe*) concerne l'emploi dans sa totalité, incluant tous les éléments connus et inconnus. En pratique, cependant, il n'est ni possible ni désirable de cerner ainsi l'univers d'un emploi. Certains aspects d'un emploi sont parfois impossibles à observer, à saisir ou à mesurer, alors que d'autres sont insignifiants en matière d'affectation du personnel. Aussi, le contenu d'un emploi est habituellement réduit aux éléments observables et non triviaux qui constituent ce que l'on appelle le domaine de contenu de l'emploi (*job content domain*). La plupart des professionnels utilisent indistinctement ces deux expressions pour désigner le domaine de contenu tel qu'il a été défini plus tôt (Cronbach, 1990).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

COMPOSANTES D'UN INSTRUMENT DE MESURE ET DE LA PRISE DE DÉCISION

Les divers éléments d'un instrument de mesure peuvent être regroupés en trois grandes **composantes** : 1) le contenu de l'instrument ou les stimuli qui sont présentés aux candidats, 2) les conditions d'application et 3) le processus d'évaluation des réponses du candidat. La figure 2.2 présente ces composantes (contour épais) et indique de quelle manière elles influencent les résultats obtenus par les candidats. Ainsi, un instrument de mesure (stimuli) appliqué à des candidats dans des conditions données permet d'obtenir des réponses qui seront ensuite soumises à un processus d'évaluation pour devenir les résultats. Les réponses et les résultats des candidats ne font pas partie de l'instrument ; ils en constituent des **résultantes** qui découlent de l'application des composantes qui les précèdent. Lorsque l'instrument de mesure sert à la prise de décision, les résultats des candidats sont alors assujettis aux règles du processus décisionnel. Les décisions qui en émanent sont une résultante de la prise de décision, alors que le processus décisionnel en est une composante.

L'organisation en cascade de ces divers éléments entraîne le principe suivant au regard de la validation basée sur le contenu : *la validité des résultats obtenus à un instrument de mesure, et des décisions qui en découlent, dépend de la représentativité de chacune des quatre composantes en rapport au domaine de l'emploi à mesurer ; l'instrument doit être tel que le type de performance exigé, le contexte dans lequel cette performance se déroule et la façon dont elle est évaluée correspondent en tout point à la réalité de cet emploi*[6]. Voyons cela plus en détail.

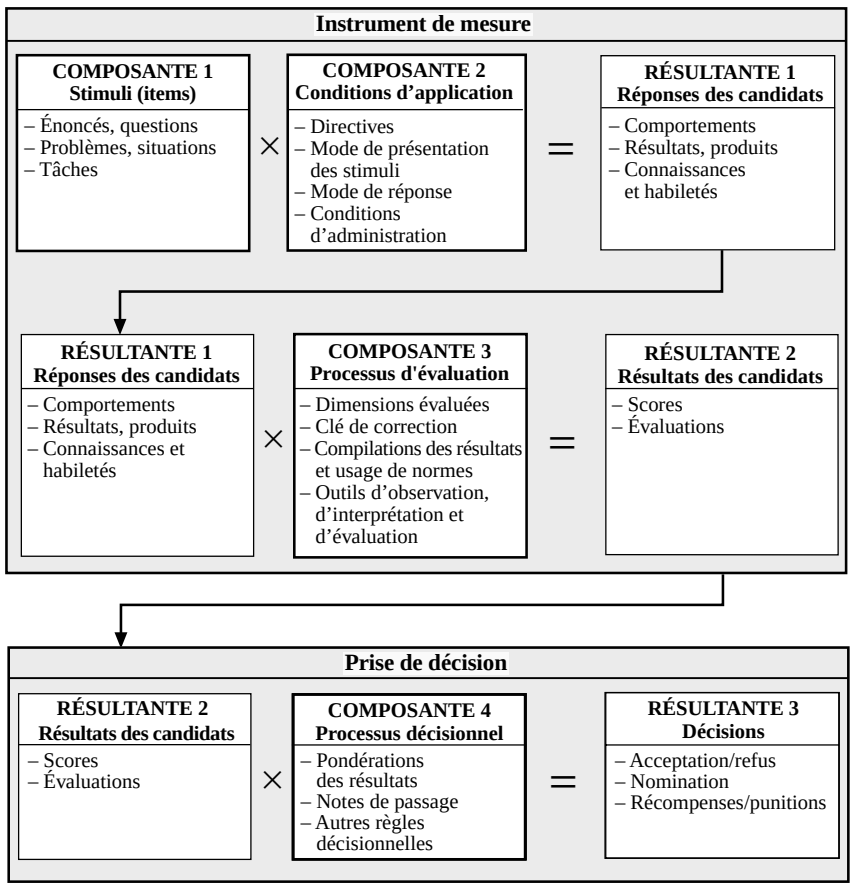
Première composante : l'instrument lui-même ou les stimuli. La première composante d'un instrument de mesure est l'instrument en lui-même, c'est-à-dire les **énoncés**, les **questions**, les **problèmes**, les **situations** ou les **tâches**, bref, les stimuli auxquels les candidats sont appelés à répondre ou à réagir. Ces stimuli, qui constituent le contenu

6. On ne retrouve pas dans la littérature une présentation aussi détaillée d'un instrument de mesure dissocié en ses composantes et ses résultantes. Par contre, Guion (1977) avait déjà exprimé que les scores obtenus à un instrument de mesure sont le fruit de plusieurs facteurs organisés en séquence. De tels scores, avait-il écrit, sont basés sur les réponses des candidats à des stimuli soigneusement standardisés et observées dans des conditions également standardisées.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 2.2
ÉLÉMENTS D'UN INSTRUMENT DE MESURE ET DE LA PRISE DE DÉCISION



de l'instrument de mesure, doivent être représentatifs du domaine à mesurer. Dans le cas d'un instrument de sélection, ils doivent refléter les situations qui sont effectivement rencontrées dans l'emploi. Pour un examen d'informatique, par exemple, les questions doivent porter sur des connaissances requises pour l'emploi concerné, les problèmes doivent être identiques, sinon similaires, à ceux qui se présentent en cours d'emploi et les tâches doivent être de même nature que celles accomplies en cours d'emploi. Ensemble, ces stimuli doivent représenter tout le domaine (ou le sous-domaine) de l'emploi concerné, et seulement ce domaine.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Pour obtenir la validité, la **forme** d'un stimulus (item) est aussi importante que son **contenu**. Cronbach (1990) illustre cette affirmation à l'aide d'un examen de connaissances dans un domaine donné où une personne, possédant l'élément de connaissance pour répondre à un énoncé de l'examen, rate l'énoncé à cause de sa formulation qu'elle ne peut comprendre. Alors cet examen n'est pas valide, parce que la performance observée à cet énoncé ne reflète pas le degré de connaissance du répondant par rapport au domaine à mesurer, mais plutôt son manque de compétence en compréhension de texte. À cause de sa forme, l'énoncé est biaisé et ne mesure pas le domaine de connaissance prévu.

Deuxième composante : les conditions d'application. La similarité avec l'emploi ne doit pas se limiter aux stimuli de l'instrument de mesure, mais doit aussi englober les conditions d'application. Cette deuxième composante de l'instrument de mesure comprend les **directives**, le **mode de présentation des stimuli** (écrit, oral, vidéo, papier, ordinateur), le **mode de réponses** utilisé (écrit, oral, comportements) et les **conditions d'administration** (durée, passation individuelle ou collective, ambiance stressante, environnement physique, etc.). Ces aspects doivent, tout comme les stimuli, bien représenter ce qui se passe dans la situation de travail afin que les réponses du candidat à l'instrument de mesure reflètent celles qu'il fournirait s'il occupait réellement cet emploi.

Par exemple, si un poste exige de réagir adéquatement sur le plan interpersonnel d'interactions avec les autres, il serait plus approprié de recourir à une mise en situation basée sur de vraies interactions avec des personnes (Tziner, Jeanrie et Cusson, 1993). Un examen écrit, où les candidats ne font que décrire leurs comportements, serait moins pertinent parce qu'il exigerait une performance différente de celle demandée dans le domaine de contenu de l'emploi. Dans cet exemple, c'est le mode de présentation des stimuli et le mode réponse qui sont en cause. Cela n'implique pas qu'un tel examen écrit ne puisse être valide en sélection, mais simplement que ces aspects ne reflètent pas le domaine à mesurer, et qu'ils peuvent altérer la validité des résultats obtenus par rapport à la performance réelle en pareille situation de travail.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Il en est de même pour les directives et les conditions d'administration de l'instrument de mesure. Idéalement, les candidats devraient être soumis aux mêmes facteurs temporels, spatiaux et psychologiques qui prévalent dans le contexte réel de travail. Par exemple, un instrument de mesure mettant l'accent sur la vitesse d'exécution n'est justifié que si la vitesse est une dimension importante de l'emploi (Cronbach, 1970).

Une première résultante : les réponses des candidats. Plusieurs éléments peuvent être considérés à titre de réponses des candidats. Selon l'instrument de mesure et le domaine de contenu à mesurer, on peut considérer soit les **comportements** des candidats (paroles, gestes ou réponses écrites), soit les **résultats ou** les **produits** de ces comportements (une soudure qui résiste à une pression donnée ou le fait qu'un candidat a réussi à convaincre les membres de son équipe), soit les **connaissances et** les **habiletés** nécessaires à ces comportements (p. ex., connaissances du point de fusion des métaux, habiletés à la négociation). On fait l'inférence que si les stimuli et les conditions d'application recréent les aspects visés du domaine de contenu de l'emploi, les réponses des candidats seront vraisemblablement typiques de celles qu'ils fourniraient en réalité.

Troisième composante : le processus d'évaluation. Le processus d'évaluation permet de décrire et de qualifier la performance des candidats à l'instrument. Leurs réponses en matière de comportements, de résultats ou de produits, de connaissances ou d'habiletés, sont maintenant analysées, interprétées puis évaluées à la lumière du domaine de contenu à mesurer, pour former leurs résultats. Comme pour les autres composantes, le processus d'évaluation doit refléter ce qui se passe dans l'emploi. Il ne sert à rien d'avoir un contenu d'instrument de mesure et des conditions d'application représentatives de la réalité, si les réponses des candidats sont observées ou évaluées à travers un processus biaisé : les résultats ne pourront qu'en être faussés. C'est le principe bien connu qu'une chaîne ne peut pas être plus forte que son maillon le plus faible. Or, le processus d'évaluation d'un instrument de mesure est souvent plus complexe à élaborer et plus difficile à appliquer que l'instrument lui-même ; il arrive fréquemment que ce soit d'ailleurs la composante la plus faible de l'instrument.

Sans entrer dans les détails, il peut y avoir plusieurs aspects à cette composante. Les **dimensions évaluées** doivent bien cerner le domaine de contenu à mesurer. Par exemple, pour une épreuve du

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

courrier (*In-Basket*), qui est une simulation individuelle papier-crayon visant à évaluer chez des candidats la capacité de résoudre des problèmes et de prendre des décisions en gestion, les dimensions testées pourraient être les fonctions suivantes du management : la planification, l'organisation, la direction et le contrôle.

La **clé de correction**, ou le solutionnaire, doit être conçue à partir des solutions reconnues comme valables dans le domaine et dans la situation réelle de l'emploi. Par exemple, uniquement pour élaborer le solutionnaire d'une épreuve du courrier utilisé au gouvernement fédéral, un comité de 14 experts a été constitué. Cadres d'expérience, occupant des emplois similaires à ceux mesurés par l'épreuve et reconnus pour être efficaces dans leur travail, ces 14 experts, répartis en deux groupes, ont dégagé ensemble un répertoire exhaustif des réponses recherchées et ont dû s'entendre par la suite sur les points à attribuer à chacune d'elles. Ainsi, lorsqu'un candidat est soumis à cette épreuve, ses réponses sont corrigées systématiquement en fonction de ce répertoire écrit.

La **compilation des résultats** ne doit pas introduire de biais dans la pondération des éléments. Par exemple, si un score total doit être compilé et si le domaine de l'emploi comporte beaucoup de planification et peu d'organisation, de direction et de contrôle, alors le score total doit refléter cette pondération. La transformation des scores bruts à l'aide de **normes**, s'il y a lieu, doit être basée sur les résultats de personnes constituant un échantillon pertinent pour la situation en cause. Par exemple, l'épreuve du courrier ci-dessus, utilisé au gouvernement, a été administrée à 99 cadres occupant le poste en question et répartis à travers toutes les régions du Canada. Par la suite, la distribution des résultats a été compilée : histogramme, moyenne, écart type, etc. (voir chapitre 7). Cette distribution constitue les normes auxquelles seront comparés les résultats des candidats à ce type d'emploi.

Finalement, lorsque le processus d'évaluation exige du jugement de la part des correcteurs, comme avec une entrevue ou un solutionnaire qui ne fournirait que les principaux éléments de réponses à valoriser, il faut s'assurer de la validité de chacune des étapes de l'évaluation. Les **observations** recueillies doivent être représentatives de la performance du candidat fournie à l'instrument de mesure, les **interprétations** doivent être exactes et non biaisées par rapport aux

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

dimensions évaluées et, finalement, les **évaluations** doivent refléter les standards et les exigences en vigueur dans l'emploi réel. Des exemples sont présentés aux chapitres 5 et 6.

Une deuxième résultante : les résultats des candidats. Le processus d'évaluation permet de transformer les réponses des candidats à l'instrument de mesure en résultats. Ces derniers peuvent prendre la forme de **scores** quantitatifs (scores, taux de réussite, temps d'exécution, nombre d'éléments produits, notation sur une échelle en cinq points, etc.), ou d'**évaluations** qualitatives (description du fonctionnement ou de la performance du candidat, ses points forts et ses points faibles, etc.). Si les composantes de l'instrument reflètent ce qui se passe dans l'emploi, les résultats des candidats seront vraisemblablement représentatifs de leur performance dans le domaine de contenu de l'emploi concerné.

Quatrième composante : le processus décisionnel. Dans la plupart des situations, les résultats obtenus à un instrument de mesure vont servir à prendre une décision, ce qui nous amène à la quatrième composante de la séquence, soit le processus décisionnel. La **pondération** ou le poids relatif de chacun des résultats doit être conséquent avec ce qui est exigé effectivement dans l'emploi. Si l'on recourt à des **notes de passage**, leur niveau doit être raisonnable et conforme à des attentes normales d'un rendement acceptable pour ce genre d'employés (Cascio, Alexander et Barret, 1988 ; EEOC, 1978).

Une troisième résultante : les décisions. Les décisions peuvent être de diverses natures. En contexte de sélection et de promotion, la décision porte essentiellement sur l'**acceptation** ou le **refus** d'une candidature pour un poste donné ; certains préfèrent utiliser le terme de **nomination**. En contexte d'évaluation du rendement, il s'agira généralement soit d'accorder une **récompense** (feed-back positif, augmentation de salaire, enrichissement du travail, etc.), soit d'infliger une **punition** (réprimandes verbales, mesures disciplinaires, contrôles accrus, etc.).

VALIDITÉ ET REPRÉSENTATIVITÉ DES COMPOSANTES

La validité porte sur les résultats, mais la validation repose sur les composantes. On se rappelle que la **validité** est une qualité qui ne s'applique pas à l'instrument de mesure lui-même, mais aux résultats

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

obtenus, ou, plus précisément, à la justesse des inférences tirées des résultats. Pourtant, dans sa quête de validité, le spécialiste ne peut pas exercer directement de contrôle sur les résultats des candidats à l'instrument, pas plus d'ailleurs que sur les autres résultantes. En revanche, il peut agir sur les composantes de l'instrument de mesure. S'il adopte une stratégie de validation basée sur le contenu, il peut faire en sorte que les stimuli, les conditions d'application ou le processus d'observation et d'évaluation reflètent le domaine de l'emploi à mesurer. Comme chaque résultante découle des composantes qui la précèdent, si ces composantes sont représentatives du domaine à mesurer, les résultantes devraient, elles aussi, être représentatives. Dans le cadre d'une stratégie de **validation** basée sur le contenu, la démonstration de la validité des résultats requiert donc l'analyse des composantes de l'instrument de mesure[7].

QU'ADVIENT-IL SI DES COMPOSANTES NE REFLÈTENT PAS LE DOMAINE DE L'EMPLOI ?

Si les composantes diffèrent de la réalité à représenter, qu'advient-il ? Faut-il rejeter d'emblée la validité des résultats obtenus à l'instrument de mesure ?

Lors d'un concours pour une promotion, une organisation recourt à un jeu de rôle pour simuler une rencontre d'évaluation entre un superviseur et son employé au sujet de son rendement. Malgré le soin apporté aux différentes composantes de la simulation, quelques

7. On retrouve un raisonnement semblable dans les propos de Messick (1975, cité dans Society for Industrial and Organizational Psychology, 1987), à savoir que toute démonstration de la valeur des résultats doit être précédée par des inférences à propos du contenu et de la méthode de construction de l'instrument de mesure. D'autres autorités, simplement en se basant sur le rationnel de la représentativité par rapport au domaine de contenu à mesurer, avaient déjà justifié l'inclusion des composantes dans la validation reliée au contenu (American Educational Research Association *et al.*, 1985 ; Equal Employment Opportunity Commission *et al.*, 1978). De toute façon, il ne peut en être autrement. On ne peut pas démontrer la représentativité des résultats d'un candidat à un instrument de mesure par rapport au domaine de contenu visé, car on ne connaît pas la performance du candidat dans le domaine en question. Si cette performance était connue, l'étude de la relation entre les résultats à l'instrument de mesure et la performance réelle serait plutôt une validation basée sur la relation avec une autre variable.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

candidats contestent leurs résultats. Ils allèguent que la simulation s'est déroulée dans un local différent de l'endroit où ils travaillent habituellement et que cela a nui à leur performance : le local est plus petit, il y a moins de fenêtres, la température y est plus élevée et les meubles n'y sont pas disposés de la même manière. Manifestement, les conditions physiques d'administration sont légèrement différentes de la réalité de l'emploi. Est-ce que cela remet pour autant en question la validité reliée au contenu de cet instrument de mesure ?

Il faut examiner les composantes de l'instrument en fonction de leur efficacité à fournir un échantillon représentatif de résultats. Il s'agit d'établir *jusqu'à quel point les divergences entre les composantes de l'instrument et la réalité de l'emploi sont de nature à entraver la mesure du domaine de contenu ciblé*. Par conséquent, un écart de ressemblance entre un instrument de mesure et le domaine de contenu de l'emploi n'altère pas automatiquement la validité, si cette carence est superficielle et ne met pas en cause les caractéristiques (*constructs*) psychologiques sous-jacentes aux réponses fournies par les candidats, objets réels de la validité (*Standards*, 1999). Réciproquement, il faut se méfier d'une similitude entre l'instrument de mesure et le domaine de contenu qui ne serait que superficielle et qui ferait en sorte que l'instrument exige des habiletés différentes de celles requises par la performance en emploi.

Dans l'exemple concernant la simulation entre un superviseur et son employé, remettre en question la validité de cet instrument sur la base de son contenu exigerait de démontrer que les différences au regard de la taille du local, des fenêtres, de la température et de la disposition du mobilier ont biaisé substantiellement la performance du candidat par rapport à une situation réelle de travail. A priori, ces différences n'ont probablement pas affecté la validité. Mais on ne sait jamais ce qu'un expert peut invoquer comme arguments et encore moins ce qu'un magistrat peut en retenir.

ANALYSE DES COMPOSANTES D'UNE ÉPREUVE DU COURRIER

Voyons un autre exemple vécu ; il s'agit d'une épreuve du courrier. Dans ce genre de test, le candidat est appelé à jouer le rôle d'un gestionnaire dans une organisation simulée et, à ce titre, à résoudre un certain nombre de problèmes qui lui sont soumis. Cette épreuve comporte 25 problèmes (items) servant à mesurer 6 habiletés de gestion

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

(la capacité d'identifier et de résoudre des problèmes, de consulter des employés et de les faire participer, d'établir un calendrier et un ordre de priorité, etc.).

Les stimuli (items). Pour appliquer une démarche de validation basée sur le contenu, il faut d'abord établir la représentativité des stimuli faisant partie de cette épreuve, à savoir les 25 problèmes soumis au candidat et les paramètres de l'organisation qui est simulée. Par exemple, est-ce que chacun des problèmes est réaliste et peut survenir dans le cadre de cet emploi ? Est-ce que l'ensemble des problèmes reflètent les six habiletés de gestion nécessaires à cet emploi ? Une fois cette démonstration complétée, il faut aussi considérer les autres composantes de l'instrument, comme les conditions d'application et le processus d'évaluation.

Les conditions d'application. En ce qui concerne le **mode de présentation des stimuli** et le **mode réponse**, l'épreuve du courrier est un test individuel et écrit, alors que dans la réalité, un gestionnaire est souvent avec d'autres et communique verbalement. Il faut se demander si ces différences au regard de la forme peuvent entraîner un biais dans les réponses fournies par le candidat. Lorsque le domaine de contenu visé est défini par des comportements et des habiletés interpersonnelles comme on en rencontre en emploi (négociation, persuasion, leadership, etc.), la validité de cette épreuve pourrait être mise en doute. En revanche, si le domaine de contenu vise des connaissances de gestion et des capacités d'analyse de problèmes (planification, organisation, jugement, etc.), un tel test individuel et papier-crayon pourrait être jugé approprié.

Poursuivons l'analyse avec les **directives** et les **conditions d'administration**. Une épreuve du courrier est une simulation de la réalité et non la réalité elle-même ; le candidat se retrouve donc dans un contexte qui n'est pas celui de l'emploi réel. Par exemple, le candidat dispose de une à trois heures pour résoudre les quelque 10 à 25 problèmes de cette épreuve, sans avoir de vue d'ensemble de la situation, sans connaître l'historique de chacun de ces problèmes ni avoir la chance de tâter le terrain avant d'opter pour une solution en particulier. Dans la réalité, un gestionnaire n'a pas à résoudre autant de problèmes de cette importance dans un laps de temps si court ; de plus, il a souvent plus d'informations sur l'organisation, sa culture, son climat, son histoire, ou il a au moins la possibilité d'en

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

savoir plus s'il le désire et d'en discuter avec son équipe. Cette disparité ne pose pas de difficultés insurmontables, à condition d'en tenir compte lors de la correction et de l'évaluation des réponses.

Pour ce qui est des **conditions psychologiques**, les différences peuvent être très grandes, surtout au regard des enjeux personnels. Dans une simulation comme une épreuve du courrier, le candidat indique ce qu'il ferait pour résoudre les divers problèmes soumis. Ce que le candidat dit qu'il ferait au test ne correspond pas nécessairement à ce qu'il ferait dans une situation réelle de travail. En effet, dans un contexte de sélection ou de promotion, les candidats sont très probablement désireux d'obtenir le poste dont il est question. Par conséquent, ils ont davantage tendance à donner des réponses plus acceptables par la société en général (Cronbach, 1990) et par l'organisation en particulier ; certains iront même jusqu'à falsifier leurs réponses.

Or, ces problèmes de désirabilité sociale et de falsification des réponses surviennent dans les simulations lorsque les énoncés ou les problèmes soumis au candidat sous-tendent des valeurs ou un jugement moral (Hunter, 1981). Il est alors important d'analyser l'impact des conditions psychologiques artificielles sur les réponses du candidat. Dans l'épreuve du courrier, le candidat sera vraisemblablement plus porté à trouver la meilleure solution possible, du point de vue des principes reconnus d'une bonne gestion ; peut-être sera-t-il même porté à répondre en fonction de ce qu'il pense être désiré par l'employeur. Sa performance dépendra beaucoup de ses capacités à trouver ainsi les réponses « désirables » à partir de ses connaissances de base en gestion, de son sens de l'analyse et de son jugement.

Ainsi, devant un problème qui exigerait une décision ferme et beaucoup de doigté de la part du gestionnaire (p. ex., intervenir auprès d'un employé qui a des problèmes d'alcool), le candidat répondra probablement qu'il prendra ses responsabilités comme le poste l'exige. Mais qu'en serait-il dans une situation réelle de travail, où cet employé serait un de ses amis intimes, ou un de ses collaborateurs privilégiés, ou une personne envers qui il est redevable ? Sa décision pourrait alors être influencée par sa loyauté à l'organisation, son sens de l'éthique ou son sens des responsabilités. Il va de soi que ce que le candidat dit qu'il ferait n'est pas nécessairement le comportement réel qu'il adopterait dans une situation de travail.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Par conséquent, la validité de cette épreuve papier-crayon dépendra de ce qu'il faut évaluer. Si le domaine à mesurer ne concerne que la connaissance des principes de base en gestion, alors les problèmes de désirabilité sociale et de falsification ne se posent pas vraiment : un candidat ne peut pas faire semblant de connaître les bonnes solutions. En revanche, si le domaine à mesurer vise les comportements futurs du candidat, les solutions qu'il appliquera vraiment une fois dans l'emploi, alors la validité de cette épreuve est sujette à discussion.

Évidemment, les problèmes de désirabilité sociale et de falsification peuvent aussi se poser pour les items dont l'application de la solution exige des comportements et des traits de personnalité qui sont hors du contrôle du candidat. Par exemple, un candidat peut dire qu'il saura déléguer ou faire appel au travail d'équipe, mais être incapable en réalité de faire confiance aux autres ou de travailler en équipe. Il en est de même pour la motivation. À partir des seules réponses du candidat, il est très difficile de prévoir la motivation dont ce dernier fara preuve dans l'exécution de sa décision ou l'application de sa solution. Encore une fois, le dire et le faire sont deux choses distinctes.

Le processus d'évaluation. Finalement, il faut aussi faire l'analyse du processus d'évaluation. Par exemple, dans le cas d'une autre épreuve du courrier, on a pu constater que tout le processus de correction reposait en grande partie sur le jugement des correcteurs : 1) définitions ambiguës des dimensions évaluées et portant à interprétation, 2) absence d'une clé de correction proposant les éléments de réponse attendus et 3) échelle de cotation en sept points dont les différents échelons sont subjectifs et imprécis. Conséquemment, la validité des scores obtenus par les candidats dépendait des correcteurs. Il a donc fallu procéder à l'analyse de l'expertise des correcteurs dans le domaine de contenu à évaluer, car cette expertise est indispensable pour garantir la validité des résultats des candidats.

Résumé. Dans le cadre d'une validation basée sur le contenu, si les composantes d'un instrument de mesure ne reflètent pas exactement les aspects du travail visé, il faut établir jusqu'à quel point cette divergence peut affecter les réponses et les résultats des candidats et mettre ainsi en péril la représentativité de ces éléments par rapport au domaine de contenu à mesurer.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

La prise de décision. Peut-on étendre le même raisonnement à la prise de décision ? Malgré l'opinion de certains spécialistes, les cours de justice reconnaissent que l'approche de validation basée sur le contenu s'applique, par exemple, à la fixation d'une note de passage (Cascio, Alexander et Barrett, 1988) ou à la pondération des résultats[8]. Que les décisions soient incluses ou non sous le vocable « validation basée sur le contenu », cela ne change rien au principe. La pondération des scores, la note de passage et les règles décisionnelles doivent pouvoir se justifier par rapport aux exigences de l'emploi.

Par exemple, la note de passage à l'épreuve du courrier avait été fixée à 30 sur 60 pour être nommé à un poste de supervision de premier niveau. Or, il a été démontré que près de 60 % des superviseurs déjà en place ne parvenaient pas à atteindre un tel rendement à cette épreuve. Il a donc été possible de convaincre le tribunal que la note de passage ne reflétait pas les exigences courantes de cet emploi.

CONDITIONS D'APPLICATION DE LA VALIDATION BASÉE SUR LE CONTENU

La validation basée sur le contenu ne s'applique pas à tous les instruments en toute situation. Parfois elle n'est pas requise, parfois elle n'est pas applicable.

Inférences descriptives. Les situations qui requièrent la validation basée sur le contenu sont celles où l'instrument sert à formuler des inférences **descriptives**, c'est-à-dire à inférer ce qui est vraiment mesuré par les résultats obtenus (p. ex., déterminer le niveau de connaissances chez un candidat en le soumettant à un examen portant sur ces connaissances, mesurer les habiletés à la négociation par une entrevue situationnelle ou apprécier le rendement d'un employé à l'aide d'une grille d'évaluation appliquée par le supérieur hiérarchique). La validité d'un instrument utilisé dans ce contexte sera assurée si les résultats mesurent bien la variable visée par cet instrument, et rien d'autre. Toutefois, la validation reliée au contenu n'est pas suffisante, ni même parfois requise, dans la plupart des cas

8. Voir la décision de Pierre Baillie, dans l'affaire Laquerre, Breault, Sinh et Ducharme contre Revenu Canada (96-NAR-00153), concernant la pondération des résultats.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

où l'instrument sert à formuler des inférences **relationnelles**, pour connaître les résultats à une variable qui n'est pas mesurée directement par l'instrument mais qui lui est reliée (p. ex., vouloir prédire le rendement au travail à partir des résultats à un test d'aptitude ou à une entrevue). La validation basée sur la relation avec d'autres variables sied mieux à ces situations.

Niveau de raisonnement peu élevé. Dans certains cas, la validation basée sur le contenu serait requise, mais elle est difficile à appliquer. Par exemple, comment un expert peut-il établir directement, par déduction logique et sans risque de se tromper, qu'un score à un test d'aptitudes mentales est de même nature que le type d'opérations mentales nécessaires pour accomplir des tâches d'analyse budgétaire ? Ou bien qu'un score à un inventaire de personnalité, mesurant la confiance en soi, représente vraiment un attribut pertinent pour un représentant commercial ? Comment établir une telle preuve par simple logique, par pure déduction, et sans se tromper, alors que les aptitudes mentales et la confiance en soi sont des dimensions abstraites ? Dans pareilles situations, relier le contenu de l'instrument de mesure aux divers éléments de l'emploi exige un certain niveau de raisonnement de la part de l'expert, et la valeur de son jugement dépendra de ses capacités en cette matière. *Plus le niveau de raisonnement requis sera élevé, plus la démonstration de la validité basée sur le contenu sera difficile et risquée* (Gatewood et Feild, 1998). On pourra présumer que ces aptitudes ou que ces traits de personnalité sous-jacents sont nécessaires à l'emploi, mais on ne pourra pas le démontrer par une simple confrontation du contenu de l'instrument et de celui de l'emploi (Arvey et Faley, 1988).

La situation idéale. Idéalement, l'application de la validation basée sur le contenu se fera dans une situation où les raisonnements sont gardés au niveau le plus bas, c'est-à-dire que l'on peut voir directement, par simple observation et presque sans raisonnement, si les composantes de l'instrument de mesure reflètent bien le contenu de l'emploi. La situation est idéale lorsque l'instrument de mesure est identique à l'emploi. Ce serait le cas d'un instrument dont les stimuli seraient les mêmes que les tâches de l'emploi réel, qui serait administré dans le même contexte que celui de l'emploi et dont les réponses seraient corrigées de la même façon qu'on évalue la performance en emploi (p. ex., conduire un camion en situation réelle, résoudre un vrai problème avec le même logiciel qu'en emploi).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Éléments concrets et observables du domaine de contenu. Aussitôt que l'on s'éloigne de cette situation idéale, que l'instrument de mesure n'est plus identique à l'emploi mais en constitue plutôt une approximation, la validation basée sur le contenu ne peut se faire sans raisonnement, sans inférence. Comme il est rarement possible qu'un instrument de mesure soit complètement identique à l'emploi, la meilleure façon de limiter le niveau de raisonnement est de réserver la stratégie de validation basée sur le contenu aux instruments qui mesurent des caractéristiques concrètes et semblables à celles observables dans l'emploi visé, telles que 1) les comportements, 2) les résultats tangibles produits par ces comportements ou 3) les connaissances et les habiletés nécessaires à ces comportements (faire fonctionner un micro-ordinateur, rédiger un discours, souder à l'arc électrique, produire des états financiers, obtenir un consensus, maîtriser les connaissances concernant la fusion des métaux, etc.). De façon générale, il est reconnu que cette stratégie de validation est particulièrement appropriée lorsqu'il s'agit de mesurer ce qu'une personne a appris à travers une formation ou par l'expérience (Cronbach, 1970 ; Nunnally et Bernstein, 1994)[9]. Cependant, si les instruments mesurent des caractéristiques plus abstraites et moins faciles à observer directement dans l'emploi (leadership, contrôle de ses émotions, créativité, esprit de synthèse, etc.), il est recommandé de faire appel à d'autres stratégies.

Résumé. Gatewood et Feild (1998) proposent huit facteurs à considérer pour évaluer si les conditions sont propices à une démarche de validation basée sur le contenu. Ces facteurs, pour la plupart, reprennent les mêmes conditions présentées plus haut. Ils peuvent se résumer aux trois conditions suivantes : 1) une situation où les instruments sont utilisés pour faire des inférences descriptives par rapport au contenu même de l'emploi, 2) le niveau de raisonnement des experts est maintenu le plus bas et le plus direct possible et 3) le contenu mesuré est constitué d'habiletés et de connaissances définies par des éléments concrets et observables de l'emploi. Ces aspects seront analysés en profondeur aux chapitres 5 et 6.

9. En anglais, on appelle « *achievement tests* », que l'on pourrait traduire par « tests de rendement », tout instrument qui mesure ce qu'une personne a ainsi appris par une formation ou par l'expérience (Nunnally et Bernstein, 1994).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

C H A P I T R E

VALIDATION BASÉE SUR LA RELATION AVEC D'AUTRES VARIABLES

Dans un contexte de sélection, de promotion ou de placement, les instruments de mesure sont utilisés pour évaluer si un candidat a les capacités, les compétences ou toute autre qualité nécessaire pour remplir un poste donné. La validation basée sur la relation avec d'autres variables vérifie, auprès d'un échantillon de candidats, jusqu'à quel point les résultats obtenus à un instrument de mesure sont reliés à d'autres variables extérieures à cet instrument, notamment au rendement au travail. Ce chapitre explique comment mener une telle étude de validation et comment juger si la valeur du coefficient de validité obtenue est suffisante. En outre, on y traite de la mesure du rendement au travail et des divers critères pouvant être utilisés. Les principaux résultats des méta-analyses et leurs formidables impacts sur la gestion des ressources humaines font également partie de l'exposé.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Plan du chapitre. Le chapitre comprend huit sections. La première situe la problématique de la validation dans le contexte de la sélection du personnel. En gestion des ressources humaines, la stratégie de validation basée sur la relation avec d'autres variables est presque toujours traitée par rapport à la sélection du personnel. Pour simplifier la présentation, il en sera de même dans ce chapitre. Le lecteur n'aura ensuite aucune difficulté à généraliser l'application des concepts présentés en contexte de promotion du personnel, de placement ou autre. La deuxième section présente le processus général d'une étude de validation basée sur la relation avec d'autres variables et les six sections suivantes abordent une à une les différentes étapes du processus. Le contenu de ce chapitre est probablement le plus complexe de cet ouvrage. Si des concepts plus techniques sont exposés, c'est dans le but très pratique de pouvoir comprendre, améliorer, défendre et remettre en question les méthodes de sélection employées dans l'organisation.

RATIONNEL DE LA SÉLECTION DU PERSONNEL

Que signifie la validité dans le cas d'un instrument de sélection ? Pour répondre à cette question, il faut connaître le rôle que jouent les instruments de mesure utilisés en sélection et ce qu'on pourrait appeler le rationnel de la sélection du personnel.

La sélection du personnel repose sur la prémisse que **les personnes sont différentes** et que plusieurs de ces différences vont subsister chez ces personnes même après avoir suivi la même formation ou avoir vécu une expérience semblable. Ainsi, il y a des personnes qui apprennent plus rapidement que d'autres, qui perdent facilement leur calme, qui sont passionnées pour tout ce qui touche la technologie, etc. Cela ne signifie pas que ces différences sont fixées de façon permanente, que ce soit par l'hérédité ou par des expériences précoces. Cette prémisse implique simplement que ces différences existent et sont raisonnablement stables chez la majorité des adultes, suffisamment du moins pour être observables sur une certaine période de temps (Guion, 1998). Ces différences entre les personnes entraînent à leur tour des **différences individuelles dans le rendement** au travail.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Or, plus ces différences individuelles de rendement sont grandes, plus il vaut la peine d'investir dans le choix des candidats afin de sélectionner ceux dont le potentiel de rendement est élevé.

Des différences individuelles de rendement substantielles et précieuses. Cook (1988) consacre un chapitre entier aux différences de rendement entre employés et à la valeur économique de ces différences. Il conclut que les écarts de productivité entre les personnes sont substantiels et que les conséquences sur l'efficacité et la rentabilité d'une organisation sont énormes. Utilisant le coefficient de variation comme indice de variabilité de la performance des employés (soit l'écart type du rendement, divisé par le rendement moyen et multiplié par 100), Schmidt et Hunter (1983) rapportent un coefficient moyen de variabilité de 21,5 % dans le cas de plusieurs emplois manuels de production. En termes plus simples, les employés les plus productifs (15 % supérieurs) ont en moyenne un rendement de 21,5 % supérieur à la moyenne. Si l'on compare les employés les plus efficaces aux moins efficaces (15 % inférieurs), les premiers ont un rendement d'environ la moitié plus élevé que les derniers. Toutefois, ils ont noté que les différences de rendement diminuent sensiblement lorsque les employés sont rémunérés à la pièce. Le coefficient de variation passe à 15 %, soit tout de même un écart de rendement d'environ un tiers entre les plus productifs et les moins productifs. Fait à souligner, l'augmentation de la quantité ne semble pas se faire au détriment de la qualité.

Ces écarts de rendement ont tendance à s'accentuer considérablement lorsque l'emploi devient plus complexe. En effet, Hunter, Schmidt et Judiesch (1990) ont obtenu un coefficient de variabilité de 19 % pour les emplois les moins complexes (ouvriers non spécialisés ou semi-spécialisés), ce qui est assez proche de celui de 21,5 % rapporté par Schmidt et Hunter (1983) pour des emplois manuels de production. Cependant, le coefficient est de 32 % pour les emplois moyennement complexes (ouvriers spécialisés, techniciens, superviseurs de premier niveau et cadres inférieurs) et de 48 % pour les emplois les plus complexes (cadres intermédiaires et supérieurs, professionnels et certains postes techniques de haut niveau). Ces coefficients indiquent des différences individuelles de rendement qui sont

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

très grandes. Dans le cas des emplois les plus complexes, par exemple, les employés les plus efficaces (15 % supérieurs) ont une production qui équivaut à une fois et demie (soit 1,48) celle des employés moyens et à près de trois fois (2,85) celle de leurs collègues les moins efficaces (15 % inférieurs)[1].

Et que vaut, sur le plan monétaire, une différence de rendement entre deux employés ? En moyenne, 40 % du salaire de l'employé pour chaque variation de rendement équivalente à une fois le coefficient de variation. Telle est l'estimation la plus conservatrice fournie par les spécialistes (Boudreau, 1991 ; Schmidt et Hunter, 1998 ; Schmidt *et al.*, 1986). Ces chiffres sont de nature à faire réfléchir sur l'importance pour l'organisation de bien choisir ses employés.

Même si tous les emplois ne se prêtent pas aisément à ce genre de quantification du rendement, les retombées économiques n'en demeurent pas moins réelles ; elles ne sont tout simplement pas quantifiées précisément, comme c'est souvent le cas en pratique. La mesure des retombées économiques des décisions en matière de personnel est un aspect important du concept d'utilité présenté plus tôt (voir chapitre 1). Si les différences individuelles de rendement sont moins importantes pour un poste donné, comme pour un emploi routinier et très encadré, il est peu rentable, en théorie, de développer un système coûteux de sélection dans une telle situation. Dans la pratique, cependant, il est rare de trouver une situation où les différences de rendement et leurs conséquences pour l'organisation sont si faibles qu'elles ne requièrent pas le recours à un processus de sélection rigoureux.

Prédire le rendement pour mieux choisir ou faire des inférences relationnelles. Étant donné l'ampleur des différences individuelles de rendement, il importe pour l'organisation de tenter de choisir les candidats dont le rendement sera le plus élevé ou, à la limite, satisfera aux exigences normales de l'emploi. Tel est l'objectif du processus de sélection. Mais il y a un problème : le rendement des candidats

1. Si l'on se réfère aux proportions de la courbe normale, environ 15 % des personnes se situent au-delà d'un écart type supérieur à la moyenne.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

demeure inconnu, puisqu'ils n'ont pas encore été engagés. La **solution idéale** à ce problème consisterait à observer directement le rendement des candidats dans la situation réelle de travail. Il suffirait d'engager un certain nombre de candidats pour une période d'essai, au terme de laquelle leur rendement serait évalué et ensuite on ne retiendrait que les candidats satisfaisants. Mais cette solution comporte habituellement trop d'inconvénients pratiques pour être appliquée telle quelle (Hakel, 1989) ; en voici quelques exemples : disponibilité des postes, conséquences d'un mauvais rendement pour l'organisation, coûts exorbitants et variant selon la durée de la période d'essai et le montant des salaires versés, déplacement de candidats détenant déjà un emploi dans une autre organisation, etc. Une telle période d'essai est un élément de solution, mais seulement à la dernière étape du processus de sélection.

Une **solution plus réaliste** est de tenter de prédire le rendement des candidats. Les qualités requises pour faire le travail sont d'abord déterminées, puis les candidats qui possèdent ces qualités sont choisis. Ainsi, la sélection du personnel consiste fondamentalement à prédire le rendement, à partir d'un modèle qui est le profil des qualités requises[2]. Les instruments de sélection servent alors à mesurer chez les candidats les éléments du profil.

Il est intéressant de remarquer qu'en situation de sélection les instruments de mesure sont utilisés pour faire des **inférences relationnelles**, c'est-à-dire pour connaître les résultats d'une autre variable (p. ex., le rendement au travail) qui est reliée aux scores obtenus à l'instrument, mais qui n'est pas mesurée directement par l'instrument (p. ex., un test d'aptitude ou un examen de connaissances). Dans ce contexte, on se rappelle que la stratégie de validation appropriée est celle basée sur la relation avec d'autres variables.

2. Les méthodes de sélection par simulations constituent une sorte de compromis entre les deux solutions, étant à la fois observation directe et prévision du rendement. À l'aide de mises en situation réalistes, le rendement des candidats est observé directement, mais en dehors d'une situation réelle de travail, ce qui évite bien des inconvénients. Une telle démarche est également de la prédiction de rendement, parce qu'elle est fondée sur la croyance que le rendement ainsi observé dans une situation artificielle est représentatif du rendement réel futur.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

ÉTUDE LOCALE DE VALIDATION BASÉE SUR LA RELATION AVEC D'AUTRES VARIABLES

La stratégie de validation basée sur la relation avec d'autres variables consiste à *vérifier empiriquement, auprès d'un échantillon de candidats, dans quelle mesure un instrument permet de sélectionner des candidats dont le rendement correspond effectivement à ce qui est exigé par l'emploi. On cherche à voir s'il y a un lien entre les résultats obtenus à l'instrument de sélection et le rendement au travail*. Ce qui sert à prédire, en l'occurrence, ce qui est censé être mesuré par l'instrument de sélection (capacité de raisonner, habileté à communiquer, connaissances techniques, maîtrise de soi, etc.), est appelé prédicteur. La variable à prédire, ici le rendement, est appelée **critère** (*Standards*, 1999)[3].

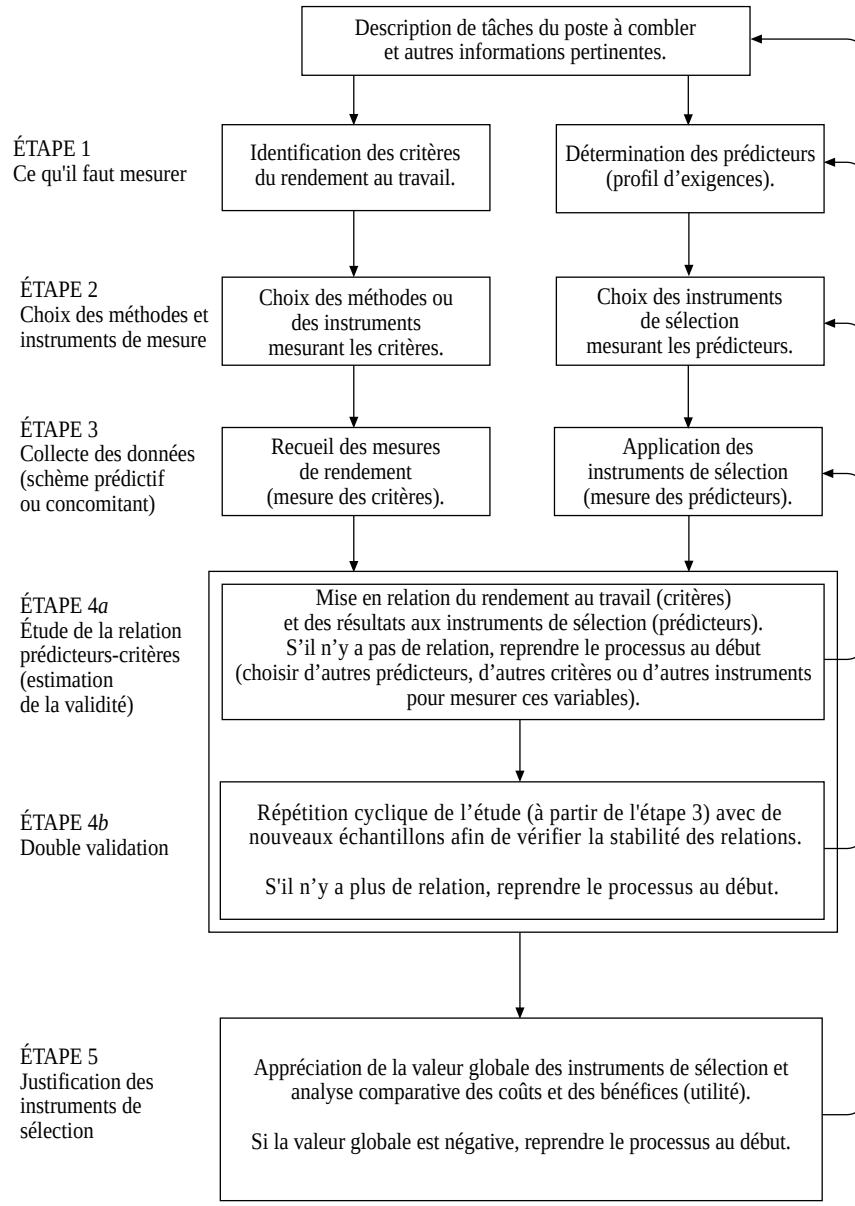
Cette stratégie de validation est expliquée à partir du cas le plus simple : un prédicteur et un critère, sans considérer d'autres variables. Ce cas permet de simplifier de nombreuses considérations conceptuelles et pratiques ; il vise à faire comprendre comment un gestionnaire peut aborder la validation de ses outils de mesure, principalement en sélection et en promotion du personnel. Quelques aspects d'un cas à plusieurs prédicteurs et plusieurs critères seront présentés par la suite.

Les étapes d'une étude locale de validation. La stratégie de validation basée sur la relation avec d'autres variables comporte différentes étapes que nous avons schématisées à la figure 3.1. Comme valider consiste essentiellement à vérifier la relation entre les résultats des candidats à l'instrument de sélection et leur rendement réel au travail, il faut pouvoir mesurer ces deux variables sur un échantillon de candidats. C'est l'objet des trois premières étapes : (étape 1) circonscrire ce qu'il faut mesurer, (étape 2) choisir les méthodes et les

3. La validation basée sur la relation avec d'autres variables vérifie si le critère de rendement peut être inféré à partir des résultats aux instruments de mesure (*Standards*, 1999). Guion (1991) rappelle que l'inférence peut être très valide pour un niveau donné du critère (p. ex., rendement faible à modéré), mais ne pas l'être pour un autre niveau du critère (p. ex., rendement très élevé).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 3.1
ÉTUDE LOCALE DE VALIDATION BASÉE SUR LA RELATION AVEC D'AUTRES VARIABLES



© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

instruments de mesure et (étape 3) procéder à l'application de ces méthodes afin de recueillir les données nécessaires à l'étude. Une fois les données recueillies, il faut procéder à l'étude de la relation entre le prédicteur et le critère (étape 4*a*) ; c'est donc à cette étape qu'est estimée la validité. L'étape 4*b* est une étape de contrôle ; elle consiste à vérifier dans quelle mesure l'estimation de validité obtenue est valable pour la situation de sélection et n'est pas le fruit d'erreurs méthodologiques.

Les étapes 1 à 4*a*, prises dans leur ensemble, constituent un processus de validation. Cette démarche, basée sur l'étude d'un seul échantillon de données, est aussi dénommée « **étude locale de validation** », appellation servant à éviter la confusion avec de nouvelles approches de validation qui intègrent de nombreux échantillons. En effet, nous verrons plus loin qu'il est possible d'estimer plus précisément la validité en agrégeant plusieurs échantillons ou plusieurs études locales. Ce nouveau type d'étude de validité sera alors appelé méta-analyse ou généralisation de la validité. Quant à la double validation, décrite à l'étape 4*b*, elle n'est qu'une autre étude locale de validation.

La figure 3.1 montre une cinquième étape qui consiste à décider s'il vaut la peine de maintenir l'usage des instruments de sélection dont la validité vient d'être étudiée. Deux justifications pourront alors être envisagées. La **première** est l'augmentation de la validité entraînée par l'application de nouveaux instruments de sélection par rapport aux instruments déjà en usage. Même si la validité d'un nouvel instrument est élevée, si elle ajoute peu à la validité déjà obtenue par les instruments en place, l'application du nouvel instrument est sans intérêt (Guion, 1998). La **deuxième** justification repose sur l'appréciation des avantages et des inconvénients et sur une analyse d'utilité, qui est essentiellement un examen des coûts et des bénéfices reliés à l'usage des nouveaux instruments de sélection (voir chapitre 1). Voyons maintenant en détail chacune de ces étapes.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

CRITÈRES DE RENDEMENT ET LEUR MESURE

IDENTIFICATION DES CRITÈRES : SPÉCIFIER LE RENDEMENT AU TRAVAIL

Le choix des critères de rendement constitue la première étape du processus de validation (voir figure 3.1, colonne supérieure gauche, étapes 1 et 2). En contexte de sélection, elle consiste à établir le rendement attendu par l'organisation, ce que l'on veut prédire ; il s'agit de définir précisément le rendement visé et ses diverses dimensions, s'il y a lieu. Les **critères de rendement** ainsi retenus sont des comportements ou des résultats reflétant les standards d'excellence à respecter pour que les objectifs de la fonction et de l'organisation soient atteints (Schneider et Schmitt, 1986). Par exemple, les critères de rendement pour un caissier dans un magasin pourraient être la vitesse d'enregistrement des articles dans la caisse, la détection des voleurs, la courtoisie manifestée avec la clientèle, etc. Pour un représentant commercial, les critères pourraient être le nombre de sollicitations, le volume de ventes, la marge de profit net, le nombre de nouveaux clients, la satisfaction de la clientèle, etc.

Dans une étude de validation, la définition des critères doit inclure les **indicateurs** tangibles de rendement plutôt que s'en tenir à des définitions plus ou moins conceptuelles (Guion, 1965, 1998). Ainsi, la qualité de l'enseignement d'un professeur d'université est une dimension relativement abstraite de son rendement, alors qu'un véritable critère pour cette dimension pourrait être le nombre de plaintes enregistrées à son égard, le nombre d'abandons chez ses étudiants ou la cote moyenne d'appréciation attribuée par ses étudiants lors de l'évaluation de son enseignement.

Les critères de rendement ne peuvent être choisis sans une analyse systématique et exhaustive du poste à combler, de la mission de l'organisation et de toute autre information relative à l'environnement interne et externe de l'organisation (voir chapitre 5, section « Analyse et description d'emploi »). La description et l'analyse des emplois constituent un champ de pratique important de la gestion des ressources humaines et occupent une place de choix dans la documentation. Plusieurs approches et techniques ont été développées au fil des ans ; la plupart de ces méthodes sont valables, chacune d'elles comportant des avantages et des inconvénients. Le choix

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

d'une méthode de description et d'analyse de tâches doit se faire judicieusement selon les objectifs poursuivis, les possibilités et les contraintes de la situation.

Choisir la méthode ou les instruments de mesure. Après avoir identifié les critères de rendement, il faut trouver une façon de les mesurer concrètement et sans biais. Dans l'exemple du caissier, la vitesse d'enregistrement des articles pourrait être mesurée à l'aide d'une caméra vidéo qui prendrait une séquence de cinq minutes à toutes les heures pendant une semaine. Il est évident que le choix des instruments qui vont servir à mesurer les critères est fortement conditionné par les critères retenus à l'étape précédente.

Comportements ou résultats ? La Society for Industrial and Organizational Psychology (1987, p. 9) et les *Standards* (1999, article 14.4) sont d'avis que tout critère de rendement doit être défini par des comportements au travail ou des résultats produits par ces comportements. Cependant, certains auteurs ne partagent pas cette orientation trop large et proposent plutôt de restreindre la performance aux seuls comportements que contrôle l'employé et qui sont pertinents aux objectifs de l'organisation : « *goal relevant actions that are under the control of the individual, regardless of whether they are cognitive, motor, psychomotor, or interpersonal* » (Campbell *et al.*, 1993, p. 40-41). La recommandation de ces derniers est fondée entre autres sur le risque de contamination, dans le cas des résultats, par des facteurs externes à l'individu. Par exemple, le volume des ventes peut être affecté par la qualité du produit, le territoire desservi ou le cycle économique. La plupart des résultats sont ainsi affectés par de multiples facteurs qui échappent à l'action de l'individu. Le présent ouvrage ne peut s'attarder à cette question. Toutefois, si des résultats sont retenus comme critères de rendement, il faudra qu'ils soient le plus possible sous le contrôle de la personne.

Critères objectifs et subjectifs. Le tableau 3.1 présente un inventaire des types de critères de rendement qu'il est possible d'utiliser à des fins de validation. À l'instar du United States Department of Labor (1970), les critères sont regroupés selon la nature objective ou subjective de leurs mesures ; un critère est dit **objectif** lorsqu'une mesure ne fait pas intervenir le jugement. Comme on peut le constater, les critères objectifs concernent habituellement des indices ou des résultats quantitatifs : nombre de pièces assemblées par unité de temps, le

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 3.1
TYPES DE CRITÈRES DE RENDEMENT

A. Critères objectifs

1. Indices de production (quantité)
Exemples : – nombre de pièces assemblées par unité de temps,
– montant des ventes,
– nombre d'étudiants supervisés ayant terminé leur thèse.

2. Indices de qualité
Exemples : – nombre d'erreurs (classement, codification, dactylo, etc.),
– nombre de plaintes (clients, subordonnés, collègues, etc.),
– coûts des reprises de travail mal exécuté.

3. Examen de connaissances et échantillon de travail
Exemples : – examen de dactylographie,
– examen de mécanique automobile,
– mise en situation (discussion de groupe, épreuve du courrier, etc.).

4. Résultats à un programme de formation
Exemples : – examen mesurant l'apprentissage d'un nouveau logiciel,
– examen mesurant la maîtrise de procédures médicales,
– examen de l'Ordre des comptables agréés.

5. Dossier de l'employé (mesures directes)
Exemples : – absences et retards,
– accidents,
– plaintes à son égard.

B. Critères subjectifs

6. Évaluation faite par le supérieur
Exemples : – mise en rang,
– notation faite sur une échelle numérique (p. ex., échelle de 1 à 5),
– échelle basée sur des comportements.

7. Évaluation faite par les subordonnés, les clients ou autres
Exemples : – notation faite sur une échelle numérique (p. ex., échelle de 1 à 5),
– recueil de divers commentaires.

8. Dossier de l'employé (mesures indirectes)
Exemples : – promotion,
– avancement salarial.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

nombre d'erreurs, le résultat à un examen de dactylographie, le nombre d'absences ou de retards, etc. Quant au critère **subjectif**, il est basé sur le jugement ou sur une évaluation faite par une personne : la notation faite par le supérieur sur une échelle numérique, une promotion, un avancement salarial, etc.

Même si les critères objectifs comportent des avantages, notamment sur le plan de la fidélité (voir chapitre 4), le rendement est extrêmement difficile à évaluer objectivement dans la plupart des cas. Par exemple, comment évaluer rigoureusement le rendement d'un professeur, d'un directeur du service à la clientèle ou d'un préposé au prêt d'équipement. En fait, l'évaluation subjective du supérieur est le critère de rendement le plus fréquent dans les études de validation (Crites, 1969, cité dans Cook, 1988 ; Lent, Aurbach et Lewin, 1971 ; Schmitt, Gooding, Noe et Kirsch, 1984). De plus, chacun des critères possède sa nature propre et mesure le rendement sous un angle qui lui est particulier ; ils ne peuvent être interchangeables (Bommer *et al.*, 1995 ; Natham et Alexander, 1988). Ainsi, plusieurs ont déjà remarqué qu'un indice de production n'est pas nécessairement relié à un indice de qualité, que l'évaluation d'un employé faite par le supérieur n'est pas forcément garante de la satisfaction exprimée par la clientèle envers ce même employé, et ainsi de suite : chaque critère comporte un aspect singulier.

La distinction entre critères objectifs et subjectifs n'est cependant pas aussi nette qu'il y paraît à première vue ; il y a souvent un élément de subjectivité même dans un critère exprimé dans des unités apparemment objectives (United States Department of Labor, 1970). Par exemple, lorsque le nombre de pièces produites par unité de temps exclut les pièces rejetées pour non-conformité aux standards, ce critère semble effectivement objectif. Pourtant, un jugement est intervenu pour établir ces standards et pour déterminer si une pièce sera acceptée ou rejetée. N'en est-il pas ainsi pour le nombre de retards au dossier de l'employé, alors qu'une personne doit décider si le retard est suffisamment grave pour l'inscrire au dossier de l'employé ? Que dire alors de l'objectivité des scores obtenus à un examen de connaissances dont la correction est basée sur le jugement du correcteur lors de l'application du solutionnaire ? Ces quelques exemples rappellent qu'il faut toujours chercher à comprendre de quoi sont constituées les mesures du critère et quelles sont les variables qui peuvent les influencer.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

QUALITÉS D'UNE BONNE MESURE DU RENDEMENT

Dans un processus de validation, le critère de rendement est aussi important que le prédicteur et, à ce titre, la mesure du critère devrait être abordée avec les mêmes exigences (Austin et Villanova, 1992 ; Binning et Barrett, 1989). Cette condition est essentielle, non seulement pour obtenir une mesure rigoureuse du rendement, mais aussi pour comprendre et interpréter correctement les résultats qui seront obtenus en ce qui a trait à la relation prédicteur-rendement. Un bon critère de rendement, qu'il mesure des comportements ou des résultats, qu'il soit objectif ou subjectif, devrait présenter plusieurs qualités.

1. Pertinence. La première qualité est la pertinence. Un critère de rendement est jugé pertinent s'il reflète des aspects significatifs de l'emploi, et seulement ces aspects. Une étude de validation basée sur la relation avec le rendement ne peut être concluante si l'on ne peut démontrer que le critère de rendement représente effectivement le domaine de l'emploi à prédire (*Standards*, 1999, article 14.4)[4]. La pertinence du critère est en fait sa **validité de contenu**. Ainsi, un aspect marginal, qui ne survient qu'exceptionnellement dans l'emploi, ne devrait pas être représenté dans le critère. Par exemple, si la personne chargée de répondre au téléphone dans une organisation ne reçoit qu'un ou deux appels en langue espagnole par année, il ne semblerait pas approprié de considérer la maîtrise de cette langue étrangère dans un critère de rendement.

La pertinence signifie également qu'il n'y a pas de **contamination** du critère par des éléments étrangers au rendement souhaité. Par exemple, le volume de ventes des représentants commerciaux peut être influencé par des inégalités des marchés qui leur sont assignés. Les critères de type subjectif basés sur le jugement d'un évaluateur sont particulièrement sujets à la contamination, notamment à cause de biais tels que l'effet de halo, les stéréotypes, le niveau d'exigence différent d'un évaluateur à l'autre, etc. Plusieurs études rapportent à ce sujet que les évaluateurs peuvent se laisser influencer, notamment par l'âge, le sexe et la race de l'évalué (Arvey et Faley, 1988).

4. La représentativité du critère correspond à l'inférence 4 présentée dans les *Standards* (1999, p. 153-154).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Il est également important que les mesures obtenues aux critères soient **indépendantes** des mesures obtenues aux prédicteurs (SIOP, 1987). Par exemple, si le critère consiste en une évaluation du rendement faite par le supérieur, il ne faudrait pas que le supérieur connaisse les résultats obtenus aux prédicteurs lors de la sélection de ses employés et qu'il risque ainsi d'être influencé dans ses évaluations du rendement. Une telle situation serait favorable à la contamination du critère. Le cas des centres d'évaluation est notable à cet égard[5]. Dans plusieurs situations, les résultats obtenus par les candidats au centre sont connus de la direction lorsque vient le temps de faire des nominations à des postes supérieurs. Par la suite, le fait d'avoir obtenu une promotion est utilisé comme critère pour valider les résultats produits par le centre d'évaluation (Gatewood et Feild, 1998).

La pertinence du critère implique la **représentativité** de l'ensemble du succès à l'emploi, ce qui signifie que le critère, ou les critères s'il y en a plusieurs, devrait être relié à l'ensemble des éléments significatifs de la fonction et dans les mêmes proportions. Si un élément important n'est pas représenté dans le critère, ou dans l'ensemble de critères, il faut redéfinir le critère ou en ajouter un autre à l'ensemble. Toutefois, selon un point de vue moins exigeant, un critère n'a pas à refléter l'ensemble des éléments de la fonction. Un bon échantillonnage pourrait suffire (Science Research Associates, 1973), à condition de pouvoir le justifier (SIOP, 1987).

La pertinence est obtenue dès la conception du critère. Nous l'avons déjà dit, le critère est établi par l'analyse détaillée des tâches du poste à pourvoir et de leur raison d'être. Au cours de cette analyse, les principaux éléments de la fonction sont définis, puis évalués selon leur contribution au bon exercice de cette fonction. En outre, il est essentiel de tenir compte des objectifs de l'organisation (Schneider et Schmitt, 1986). Il est évident que la pertinence d'un critère, même lorsqu'elle est établie de façon rationnelle, fait appel au jugement de personnes habilitées à évaluer et reflète, jusqu'à un certain point, les valeurs privilégiées par une hiérarchie.

5. Un centre d'évaluation est une méthode de sélection qui utilise plusieurs instruments de mesure, notamment les mises en situation.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

2. Logiquement relié aux prédicteurs à l'étude. Une deuxième qualité d'un critère de rendement est qu'il doit être logiquement relié aux prédicteurs à l'étude (Science Research Associates, 1973 ; United States Department of Labor, 1970 ; Wernimont et Campbell, 1968)[6]. Prenons l'exemple d'un emploi en usine dont les principales dimensions du rendement sont 1) les compétences techniques et l'apprentissage rapide des nouvelles technologies, 2) l'assiduité au travail et 3) les relations avec les autres. Pour mesurer ces éléments, on a recours aux critères suivants : 1) les résultats aux divers programmes de formation en usine, 2) le dossier de l'employé et 3) l'appréciation du superviseur. Supposons maintenant que l'on veuille valider une batterie de tests d'aptitudes cognitives utilisée en sélection de personnel dans cette organisation. Il serait plus avisé d'étudier la validité de ces tests en fonction du premier critère (« résultats aux programmes de formation »), parce que les aptitudes cognitives représentent justement la capacité d'apprentissage et de développement de nouvelles compétences. De plus, il ne faudrait pas retenir les deux autres critères pour cette étude de validation car les aptitudes cognitives ont peu à voir avec les comportements comme l'assiduité au travail et les relations avec les autres ; ces comportements dépendent plutôt de la motivation et de la personnalité de l'individu. Par conséquent, tout recours aux deux derniers critères, qu'ils soient pris isolément ou combinés dans un critère global avec la dimension « compétences techniques et apprentissage », affaiblirait la relation entre les tests d'aptitudes cognitives et de telles mesures du rendement, sous-estimant ainsi la validité réelle des tests[7].

3. Rendement individuel et sous le contrôle de l'employé. Comme le critère doit être logiquement relié aux prédicteurs et que les prédicteurs sont des caractéristiques individuelles, il serait naturel de recourir à un critère qui exprime une performance **individuelle** (Gatewood et Feild, 1998). De la même manière, le critère doit représenter des facettes du rendement soumises au **contrôle** direct de l'individu, ce qui n'est pas toujours le cas (Campbell *et al.*, 1993). Par exemple, l'équilibre budgétaire d'un directeur d'établissement public est influencé

6. Cet aspect concerne l'inférence 3 présentée dans les *Standards* (1999, p. 153-154).
7. D'une certaine manière, la validité synthétique (Lawshe, 1952, cité dans Guion, 1998) peut être considérée comme une application de ce principe.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

directement par la masse salariale, laquelle est régie par des conventions collectives négociées en amont par le ministère concerné. Ainsi, la qualité du travail effectué par une personne peut être affectée par une foule de facteurs dont le contrôle lui échappe, comme les matières premières, l'équipement mis à sa disposition, la nature de la supervision, etc. Ce sont autant de sources de contamination dont il faudra tenir compte lors de l'analyse des résultats. Toutefois, il peut y avoir des exceptions à ces règles. Par exemple, il serait justifié de recourir à un critère de rendement collectif si l'on voulait comparer plusieurs équipes de travail ; l'échantillon serait alors constitué non pas d'individus mais d'équipes. Le rendement de chaque équipe serait mis en relation avec les résultats au prédicteur pris globalement pour toute l'équipe.

4. Fidélité. Une quatrième qualité d'un critère est sa fidélité ; un chapitre complet porte sur ce concept (voir chapitre 4). Un critère est fidèle dans la mesure où il n'est pas affecté par des erreurs aléatoires : distraction d'un évaluateur en remplissant le formulaire, niveau d'exigence qui fluctue selon la période de l'année, mémoire défaillante de l'évaluateur, etc. Même si la fidélité est une qualité importante, elle ne doit jamais primer sur la pertinence (Schneider et Schmitt, 1986) ; pourtant, c'est une erreur très répandue. En effet, à vouloir mesurer ce qu'elles peuvent mesurer objectivement (fidélité) et non ce qu'elles doivent mesurer (pertinence), les grandes bureaucraties aux multiples niveaux hiérarchiques en viennent souvent à réduire le rendement de leurs employés à quelques indices objectifs mais incomplets (nombre de dossiers traités, sans égard à la qualité, montant des projets réalisés, etc.).

Il faut réaliser que le manque de pertinence d'un critère a des conséquences plus irrémédiables que le manque de fidélité. En effet, il est difficile d'éliminer après coup l'effet des biais sur les mesures, ne connaissant ni l'ampleur ni la nature exactes de ces biais. Cependant, il n'est pas essentiel que le critère soit extrêmement fidèle (SIOP, 1987). Le manque de fidélité va abaisser l'estimation de la validité observée. Il s'agit alors d'évaluer la fidélité du critère, puis de corriger la validité observée à l'aide de la formule appropriée (voir plus loin la section « Appréciation de la grandeur du coefficient de validité »).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

5. Capacité de discriminer. Un critère peut être pertinent et fidèle, et se révéler complètement inutile pour une étude de validation s'il ne permet pas de distinguer un bon employé d'un autre moins bon. Une cinquième qualité d'un critère réside donc dans sa capacité de discriminer différents niveaux de rendement (Guion, 1965, 1998 ; Schneider et Schmitt, 1986). Lors de la présentation du rationnel de la sélection, on a mentionné que la sélection, dont l'objectif est de choisir les meilleurs candidats, n'était souhaitable que dans la mesure où ces différences de rendement existaient réellement. Ainsi, un critère de rendement, lorsque tous les employés ont plus ou moins le même, est inutile dans une étude de validation : il doit y avoir de la variation. Par exemple, si tous les employés se sont absentés le nombre de jours permis par la convention collective, l'absentéisme n'est pas un critère à retenir.

De la même façon, un critère qui n'est pas sensible aux différences pourtant réelles de rendement ne peut pas servir à vérifier la validité d'un instrument de sélection. Supposons une organisation dont presque tous les employés syndiqués reçoivent la cote « satisfaisant » à leur évaluation annuelle basée sur l'appréciation du superviseur, et cela, sans égard aux différences pourtant observables de leur rendement. Une telle évaluation n'est probablement pas assez sensible aux niveaux réels de performance et il faudrait trouver un autre critère de rendement pour l'étude de validation[8]. Notons que, du point de vue de l'organisation, cette évaluation du rendement pourrait être tout à fait correcte si, par exemple, dans un contexte de valorisation du travail en équipe, la politique de l'organisation en matière d'évaluation du rendement était de ne souligner que les niveaux de performance exceptionnellement hauts ou exceptionnellement bas.

6. Pratique. Une sixième qualité d'un critère de rendement est son caractère pratique, c'est-à-dire qu'il doit pouvoir être mesuré facilement, sans interférer avec le bon fonctionnement de l'organisation. Guion (1965) rappelle, à juste titre, que l'organisation n'existe pas pour satisfaire le service du personnel. Au contraire, les pratiques

8. Cette question importante de la capacité de discriminer du critère est abordée à deux reprises lors de l'étude de la relation prédicteur-critère : une première fois sous l'angle de la variation insuffisante à la section « Vérifications préliminaires » et une deuxième fois sous l'angle de la restriction de l'étendue de l'échantillon à la section « Appréciation de l'ampleur du coefficient de validité ».

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

de gestion des ressources humaines sont là pour contribuer à l'atteinte des objectifs fondamentaux de l'organisation : profits, adaptation, survie, etc. Si la mesure du rendement exige des efforts tels que les opérations sont ralenties ou qu'une partie importante des ressources est détournée des activités productives, il faut s'interroger sur la valeur réelle d'une telle démarche. Ne perdons pas de vue que le contrôle dans une organisation doit demeurer un moyen et non une fin.

CRITÈRE SIMPLE OU MULTIPLE

Un des aspects les plus débattus par les spécialistes en validation concerne le nombre de critères de rendement et leur composition. En effet, il est admis par la plupart que le rendement au travail est un concept complexe renfermant **plusieurs dimensions** (Cascio, 1987). Ainsi, les grilles d'évaluation du rendement en usage dans les organisations incluent toujours diverses dimensions à évaluer, parmi lesquelles se retrouvent des aspects comme la quantité et la qualité du travail accompli, les relations avec les autres, la fiabilité, l'effort, le désir d'apprendre, etc. Alors, faut-il utiliser plusieurs critères, à raison d'un par dimension du rendement, ou n'en retenir qu'un seul ? Dans la littérature, on fait ordinairement référence à ces deux options en parlant respectivement de **critère multiple** et de **critère simple**. Dans le cas d'un critère simple, il y a deux possibilités : la première est un critère unique comportant la mesure d'une seule dimension (p. ex., le volume des ventes ou le nombre d'accidents) ; la deuxième est un critère composite. Il est formé de la mesure de plusieurs dimensions, regroupées ensuite pour donner une mesure globale du rendement (p. ex., évaluation moyenne du supérieur par rapport à 10 dimensions du rendement, chaque dimension étant évaluée sur une échelle en cinq points).

D'un point de vue strictement logique, si le rendement est multidimensionnel, une mesure adéquate de celui-ci devrait également être multidimensionnelle. De plus, il va de soi que chacune des dimensions du rendement, par sa nature propre, peut relever de compétences ou d'attributs distincts chez le candidat. Une dimension du rendement, comme l'effort et la persévérance, dépendra surtout des besoins et des valeurs de l'individu, alors que la qualité de son travail dépendra certainement de sa compétence technique, et ainsi de suite. Or, une étude de validation a justement pour objet d'établir empiriquement la relation entre les compétences ou les attributs de

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

l'individu et son rendement dans un emploi donné. Étant donné que chaque attribut est une caractéristique différente de la personne et que chacun de ces attributs peut être relié de façon singulière aux diverses facettes de la performance, il faut mesurer la performance sous ses diverses facettes si, dans un processus de validation, on veut étudier rigoureusement l'ensemble des relations.

Le recours au critère multiple semble donc incontournable pour respecter la rigueur scientifique, bien qu'il puisse en être autrement en pratique. Selon Cascio (1987), le choix d'un critère simple ou multiple dépend du contexte de l'étude. Si la validation est réalisée dans un contexte pratique pour appuyer les décisions de sélection, les candidats devront éventuellement être classés, du plus fort au plus faible. Un critère global peut alors suffire, car il traduit cet objectif. En revanche, si le contexte est celui d'une recherche visant la compréhension des déterminants du rendement au travail, l'emploi de plusieurs critères ou de plusieurs dimensions est essentiel.

Dimensions génériques du rendement. Il est clair que la compréhension des déterminants du rendement au travail passe par l'étude systématique des relations entre les nombreux attributs de la personne et les diverses dimensions du rendement. À cet égard, quelques auteurs ont tenté d'identifier les dimensions génériques du rendement de la performance au travail (voir chapitre 5, à la section « Transposition, sans inférence, en comportements ou en résultats » au paragraphe « Dimensions génériques du rendement »).

AUTRES CONSIDÉRATIONS RELATIVES AU CRITÈRE DE RENDEMENT

Il existe d'autres questions importantes qu'il faut envisager lors de la mesure du critère. Ainsi est-il préférable d'avoir un critère qui reflète le rendement typique observé quotidiennement ou le rendement maximum dont est capable la personne ? Vaut-il mieux un critère à court terme ou à long terme ? Comment tenir compte de la nature dynamique, changeante du rendement ? Ces questions et plusieurs autres ont attiré l'attention de plusieurs auteurs depuis bien des années (Austin et Villanova, 1992 ; Cascio, 1987 ; Guion, 1998 ; Gatewood et Feild, 1998). Paradoxalement, après des décennies de recherches dans un monde occidental si entiché d'efficacité, définir et mesurer convenablement la performance de chaque membre d'une organisation demeure encore aujourd'hui un problème irrésolu.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

PRÉDICTEURS ET LEUR MESURE

Le choix des prédicteurs et de leurs instruments de mesure, comme pour les critères, s'effectue en deux étapes (voir figure 3.1, colonne supérieure droite, étapes 1 et 2).

Identification des prédicteurs : établir le profil d'exigences. La première étape consiste à dresser le profil des exigences de l'emploi, à savoir les qualifications et autres caractéristiques individuelles des candidats. Par exemple, les prédicteurs pour un caissier dans un magasin pourrait être un diplôme d'études secondaires, deux ans d'expérience de travail auprès du public, le souci du détail, l'autonomie, un certain intérêt pour le travail routinier, l'honnêteté, etc. Pour un représentant commercial, les prédicteurs pourraient être un diplôme d'études collégiales, trois ans d'expérience dans la vente et la sollicitation, la connaissance du produit et de ses marchés, la maîtrise des logiciels *Word* et *Excel*, la confiance en soi, la persévérance, l'intérêt à faire de nombreuses heures de travail, etc. Les **prédicteurs** sont les divers éléments de formation, d'expérience, de connaissances, de compétences techniques, d'aptitudes, de traits de personnalité, d'intérêts, de valeurs et les autres caractéristiques requises pour effectuer le travail avec succès, selon les critères de rendement identifiés plus tôt. La définition des prédicteurs doit aussi inclure des **indicateurs** observables servant à les concrétiser. Ainsi, il ne suffit pas de mentionner que le candidat devra avoir de l'intérêt pour l'informatique, mais il faut rendre mesurable cette exigence grâce à des indices tangibles comme cinq ans d'expérience dans le domaine, connaissance des principales revues ou inscription à du perfectionnement. Les indicateurs doivent en outre être en nombre suffisant pour constituer un échantillon représentatif du prédicteur visé.

La détermination des prédicteurs se fait habituellement à partir de **l'analyse des tâches** du poste à pourvoir et de son environnement (voir chapitre 5, section « Analyse et description de l'emploi »). Une fois les tâches ainsi décrites en détail, leurs conditions d'exécution et leur environnement définis, il s'agit de déterminer par **déduction logique** les variables individuelles nécessaires pour accomplir adéquatement chacune de ces tâches. La définition des compétences et des qualités requises pour un emploi sera grandement facilitée par des connaissances en psychologie du comportement au travail. Un modèle intégré des principaux déterminants

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

individuels et organisationnels du rendement de l'individu au travail peut être un atout (consulter le chapitre 5, section « Transposition, avec inférence, en caractéristiques individuelles sous-jacentes »).

L'analyse et la description de l'emploi ne sont pas les seules sources d'informations pour trouver les prédicteurs. On peut emprunter une voie très **empirique** et consulter les études de validité, études locales et méta-analyses, ayant porté sur des postes similaires. Examiner de cette façon les résultats obtenus par d'autres spécialistes est de nature à fournir un éclairage objectif et réaliste sur la validité de nombreux prédicteurs. Plusieurs de ces **méta-analyses** sont mentionnées plus loin (section « Consulter les résultats des méta-analyses »). Quant aux **études locales** de validation, elles pourront être repérées au moyen d'une recherche bibliographique ; quelques-unes sont répertoriées à la section « Consulter les autres études locales de validation ».

Choisir les instruments de mesure. Une fois le profil d'exigences établi, la deuxième étape porte sur le choix des instruments pour mesurer ces prédicteurs. Les instruments de mesure les plus courants en sélection sont l'entrevue sous toutes ses formes, les examens de connaissances, les tests psychométriques d'aptitudes et de personnalité, les mises en situation et les échantillons de travail. On peut alors opter pour des **instruments existants** déjà offerts sur le marché, comme c'est le cas des tests psychométriques, ou décider de **construire ses propres instruments**, comme c'est souvent le cas pour les instruments que l'on désire davantage spécifiques au poste. Choisir un outil existant ou en construire un nouveau exige la maîtrise de nombreux concepts de mesure, dont plusieurs font partie du présent volume. De plus, le chapitre 1 présente des critères et un cadre d'analyse pouvant servir à l'évaluation des divers types d'instruments de sélection. Enfin, si l'on souhaite se faire une opinion sur un test existant, l'imposante collection des *Mental Measurement Yearbook*, éditée par The Buros Institute of Mental Measurement et présentant des analyses critiques des principaux tests sur le marché, constitue une source d'information particulièrement utile à cet égard ; cette documentation peut être consultée par l'Internet (http://www.unl.edu:80/buros/).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

COLLECTE DES DONNÉES

Une étude de validation est essentiellement une opération de vérification, qui vise à contrôler le niveau de la relation entre les prédicteurs et les critères ; cette vérification est effectuée à l'étape 4*a* du processus de validation (voir figure 3.1). Mais, auparavant, il faut avoir mesuré ces prédicteurs et ces critères. C'est l'objet de l'étape 3, au cours de laquelle la collecte des données est effectuée en appliquant à un échantillon de sujets les méthodes et instruments de mesure retenus lors des étapes précédentes. Rappelons, pour l'instant, qu'un seul critère et un seul prédicteur sont considérés.

SCHÈME PRÉDICTIF OU CONCOMITANT

Schème prédictif. Tiffin (1942, cité dans Guion, 1991) a proposé deux façons de procéder à la collecte des données. La première façon est le schème prédictif, également appelé la méthode longitudinale. Suivant ce schème, le prédicteur est mesuré dès le processus de sélection des candidats, avant qu'ils n'occupent leur poste, alors que le rendement est mesuré après leur embauche au terme d'une période passée à l'emploi (voir figure 3.2). C'est une approche longitudinale qui se déroule selon l'ordre naturel du processus de sélection habituel.

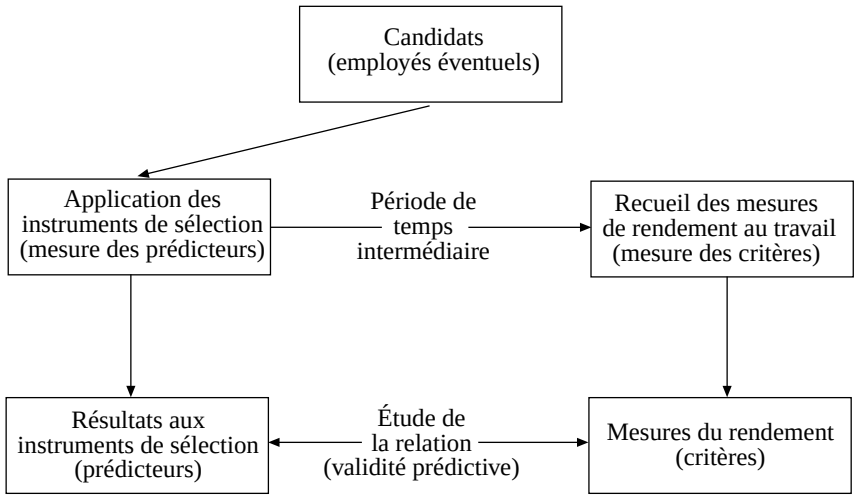
Par exemple, afin de valider un examen de dactylographie pour embaucher du personnel de secrétariat, on demande à toutes les personnes qui postulent de passer cet examen lors du processus de sélection. Par la suite, on retient les meilleurs candidats suivant la procédure de sélection habituelle[9]. Après une période d'adaptation de trois mois au cours de laquelle les candidats embauchés ont pu se familiariser avec leur nouveau travail, on procède à la mesure de leur rendement. Un autre exemple est fourni par les finissants au baccalauréat en sciences comptables qui se présenteront à l'examen de l'Ordre des comptables agréés. Pour établir si leur moyenne cumulative au baccalauréat peut prédire leur résultat à l'examen de l'Ordre, il s'agit de compiler leur moyenne cumulative (prédicteur) obtenue à

9. Lors d'une procédure de validation, l'instrument à l'étude devrait idéalement être utilisé à titre expérimental, sans tenir compte des résultats obtenus par les candidats dans la décision d'embauche. Cet aspect est abordé plus loin (voir section « Appréciation de la grandeur du coefficient de validité »).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 3.2
SCHÈME PRÉDICTIF OU MÉTHODE LONGITUDINALE



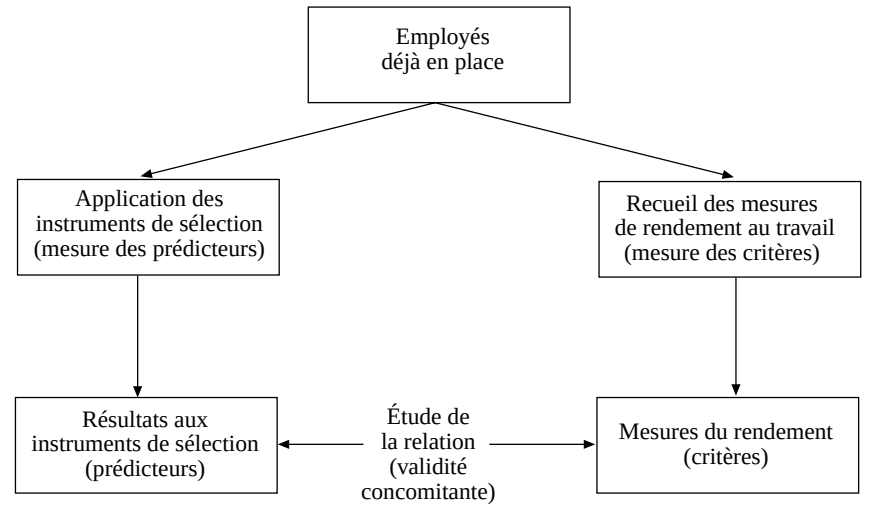
la fin de leur baccalauréat au mois de mai, puis de recueillir en décembre de la même année leur résultat à l'examen de l'Ordre (critère) qu'ils auront passé au cours de l'automne.

Schème concomitant. La deuxième façon de recueillir les données s'appelle le schème concomitant ou la méthode des employés en place. Cette fois, il n'est pas question de procéder avec des candidats qui postulent réellement un emploi ni d'attendre une certaine période après leur avoir appliqué l'instrument de mesure à valider. Il s'agit plutôt d'expérimenter l'instrument de sélection sur un groupe d'employés travaillant déjà dans l'emploi étudié. Le prédicteur et le critère sont mesurés de manière concomitante, c'est-à-dire par rapport à la même période de temps (voir figure 3.3).

Par exemple, une organisation désire implanter un épreuve de courrier (*In-Basket Test*) pour la sélection de cadres. Cependant, l'organisation veut s'assurer au préalable que cet instrument permet vraiment de découvrir les cadres les plus efficaces. Elle décide donc de mettre sur pied une étude pilote, au cours de laquelle cette épreuve sera expérimentée auprès d'un échantillon de cadres déjà à l'emploi de l'organisation. Après leur avoir expliqué la raison de cette étude,

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 3.3
SCHÈME CONCOMITANT OU MÉTHODE DES EMPLOYÉS EN PLACE



on demandera aux sujets de l'échantillon de répondre de leur mieux à l'examen, comme en situation de sélection. Parallèlement à cette expérimentation, on recueille les données sur le rendement de ces cadres.

Comparaison des schèmes prédictif et concomitant. Le schème concomitant offre l'**avantage** de faire connaître assez rapidement les résultats concernant la validité de l'instrument étudié. Il n'est pas nécessaire, comme pour le schème prédictif, d'attendre après la période d'acclimatation ni d'avoir engagé suffisamment de candidats pour constituer un échantillon valable (ce qui peut être fort long, et prendre même plusieurs années, dans maintes circonstances). Toutefois, cet avantage par rapport au schème prédictif comporte aussi plusieurs **inconvénients**, dont voici les principaux (Arvey et Faley, 1988 ; Cook, 1988 ; Tiffin et McCormick, 1958). Premièrement, il peut survenir un **biais dans l'échantillon**. Les employés en place ont tendance à constituer un groupe sélectionné différent de la population des candidats réels : certains d'entre eux ont déjà été éliminés lors de la sélection, d'autres qui n'avaient pas un rendement satisfaisant ont été amenés à quitter ou sont partis de leur propre gré, pendant que certains dont le rendement était très satisfaisant se sont peut-être vus offrir une promotion. Bref, les personnes ont tendance

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

à se déplacer vers les emplois pour lesquels elles sont le mieux qualifiées et dans lesquels elles se sentent le plus satisfaites. Par conséquent, les employés en place sont susceptibles de former un échantillon plus homogène que la population des candidats en général pour ce qui est des qualifications pertinentes au poste et du rendement. Deuxièmement, les employés en place ont probablement changé en développant des habiletés nécessaires pour leur emploi. Forts de l'expérience acquise à l'emploi étudié, les employés en place pourront être **avantagés** avec tous les instruments de sélection qui mesurent des caractéristiques se développant avec l'apprentissage. Troisièmement, le contexte d'étude pilote dans lequel évoluent les employés en place est fort différent de celui auquel sont soumis les candidats réels. En effet, dans un schème concomitant, il est probable que les employés en place manifestent **des attitudes et des comportements** différents de ceux qu'adopteraient des candidats réels (p. ex., motivation de réussir, anxiété liée au fait de ne pas obtenir l'emploi convoité ou propension à falsifier les réponses).

Étant donné que les employés déjà en place représentent une population distincte de celle des candidats et qu'ils sont testés dans des conditions différentes, Guion (1965) est d'avis que l'étude de validation concomitante ne respecte pas les principes scientifiques. Malgré cela, plusieurs chercheurs ont observé que les coefficients de validité obtenus dans les études concomitantes étaient pratiquement identiques à ceux des études prédictives (Barret, Phillips et Alexander, 1981 ; Schmitt *et al.*, 1984). Hartigan et Wigdor (1989) ont fait la même observation auprès de 755 échantillons à l'aide de la *Batterie générale de tests d'aptitudes* (BGTA). Par conséquent, on serait porté à considérer les coefficients de validité concomitante comme de bonnes estimations de la validité prédictive. Mais Guion (1991) fait remarquer, à juste titre, que l'absence de différence entre les deux formes de validité ne vaut que pour les études utilisant les tests d'aptitudes cognitives. Pour les autres types de prédicteurs, les résultats des études empiriques ne sont pas aussi convaincants. Il est donc plus prudent, surtout pour ces prédicteurs, de s'abstenir d'affirmer que la méthode de collecte des données n'a pas d'influence sur l'estimation des coefficients de validité.

La méthode prédictive semble préférable à la méthode concomitante, bien qu'elle exige plus de temps. Il est possible de recourir à la méthode concomitante pour évaluer rapidement la validité d'un outil

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

de sélection, à condition d'être conscient des biais possibles. De plus, il serait sage d'effectuer un suivi auprès des candidats sélectionnés en appliquant cette fois la méthode prédictive[10]. Cependant, le schème concomitant ne devrait jamais être retenu si les nouveaux employés sont choisis habituellement parmi des personnes inexpérimentées et si l'expérience acquise en cours d'emploi influence la performance au prédicteur (Gulliksen, 1950, cité dans Guion, 1991).

CHOIX DE L'ÉCHANTILLON

Qu'on recoure à un schème prédictif ou concomitant, la valeur des résultats d'une étude de validation repose avant tout sur l'échantillon. L'American Educational Research Association, l'American Psychological Association et le National Council on Measurement in Education soutiennent que la validité d'un instrument de mesure peut être estimée à l'aide d'une seule étude locale, mais à condition notamment que l'échantillon soit suffisamment grand, qu'il représente la population visée et que, bien entendu, le critère soit approprié (1999, article 14.3). Ainsi, l'échantillon retenu dans une étude de validation doit d'abord être suffisamment **grand**. En recherche, la taille de l'échantillon est directement reliée au concept de puissance statistique (voir chapitre 7, section « Coefficients de corrélation et de détermination »). Un grand échantillon permet de réduire les erreurs aléatoires d'échantillonnage et augmente d'autant les chances de découvrir s'il y a une relation réelle entre le prédicteur et le critère. Pour une étude de validation, Schmidt, Hunter et Urry (1976) recommandent un échantillon de 172 sujets et plus afin d'avoir une chance raisonnable de détecter une validité réelle d'ampleur moyenne ; d'autres auteurs ont publié au sujet de la taille requise de l'échantillon (Sacket et Wade, 1983 ; Raju, Edwards et Loverde, 1985, cité dans Guion, 1991). Quant à l'erreur d'échantillonnage, elle est expliquée

10. À cet égard, il est intéressant de consulter les travaux de Guion et Cranny (1982), qui présentent cinq schèmes de collecte de données en contexte de sélection dans lesquels la mesure du prédicteur précède la mesure du critère ; certains de ces schèmes sont plus recommandables que d'autres pour une étude de validation. Voir aussi l'article de Sussmann et Robertson (1986) portant sur l'analyse comparative des diverses méthodes de collecte de données lors d'une étude de validation.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

plus loin (voir les sections « Considérer les faiblesses méthodologiques », « Double validation » et « Méta-analyse, généralisation de la validité et autres méthodes de validation »).

L'échantillon doit aussi être **représentatif**. En effet, il ne sert à rien d'avoir un échantillon de plusieurs centaines de sujets si ces derniers ne reflètent pas la population cible. L'échantillon doit être représentatif de l'ensemble des candidats, pour toutes les variables présumées avoir une influence sur la relation entre le prédicteur et le critère (SIOP, 1987). Que ce soit l'âge, le sexe, la race, l'instruction, l'expérience, etc., ces variables doivent être représentées dans l'échantillon et selon les mêmes proportions que dans la population de candidats. En outre, si des candidats se sont retirés de l'étude ou en ont été exclus, il faut en donner les raisons.

ÉTUDE DE LA RELATION PRÉDICTEURS-CRITÈRES

Nous voici à l'étape où la validité peut être estimée : il s'agit d'évaluer la relation qui existe entre les résultats obtenus à l'instrument (prédicteur) et le rendement au travail (critère). Cette étude est effectuée à l'étape 4*a* du processus de validation (voir figure 3.1).

Étudier rigoureusement et systématiquement la relation entre deux variables est d'abord une affaire de statistique, domaine au sujet duquel il existe de très nombreux ouvrages. Par conséquent, il n'est pas souhaitable de reprendre dans cet ouvrage l'ensemble des concepts qui s'y rapportent. Le bref exposé au chapitre 7 suffit à éclairer ce qui suit et le lecteur intéressé à approfondir ces concepts ou devant mener lui-même une étude de validation importante est prié de consulter des ouvrages spécialisés de psychométrie et de statistique.

NIVEAUX DE MESURE DES PRÉDICTEURS ET DES CRITÈRES

En statistique, la façon d'analyser la relation entre deux variables doit tenir compte du niveau d'échelle de mesure des variables à étudier. Seuls les trois premiers niveaux d'échelles sont habituellement considérés : les mesures nominales, ordinales et d'intervalle. Le quatrième niveau, qui est l'échelle de rapport, est assimilé à l'échelle d'intervalle lors de l'application de la plupart des méthodes statistiques (voir chapitre 7, section « Échelles de mesure ». Dans la

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

pratique, on retrouve des prédicteurs et des critères dont le niveau de mesure peut être nominal, ordinal ou d'intervalle. Par exemple, être bilingue, avoir un permis de conduire, être citoyen canadien ou détenir un diplôme en informatique sont des **prédicteurs** de niveau nominal, alors qu'être le premier aux examens de l'Ordre des comptables agréés ou faire partie des trois meilleures équipes au concours interuniversitaire des Jeux du commerce sont des prédicteurs de niveau ordinal. La moyenne cumulative au baccalauréat, le nombre de fautes d'orthographe dans une dictée d'une page ou le score sur une échelle en cinq points obtenu à la suite d'une entrevue seraient plutôt des exemples de prédicteurs de niveau d'intervalle. Quant aux **critères**, avoir réussi un programme d'entraînement à la tâche, avoir complété avec succès sa période de probation ou avoir été promu constituent des exemples de critères de niveau nominal, le niveau ordinal visant plutôt des critères comme faire partie des trois meilleurs vendeurs du mois ou être le professeur le plus apprécié des étudiants. Des exemples de critères de niveau d'intervalle seraient le volume de ventes, le nombre d'unités produites ou certaines formes d'évaluation du rendement exprimée sur une échelle numérique.

Combinaisons prédicteur-critère. La combinaison des trois niveaux de mesure pour le prédicteur et pour le critère donne lieu, théoriquement, à neuf situations différentes, illustrées par la figure 3.4. À la première ligne, on retrouve un critère de niveau nominal combiné à un prédicteur nominal (combinaison 1), à un prédicteur ordinal (combinaison 2) ou à un prédicteur d'intervalle (combinaison 3). À la deuxième ligne, on retrouve les mêmes combinaisons avec, cette fois-ci, un critère dont l'échelle est ordinale (combinaisons 4, 5 et 6), alors qu'à la troisième ligne, le critère est de niveau d'intervalle (combinaisons 7, 8 et 9).

En contexte réel de gestion des ressources humaines, il est peu fréquent de rencontrer des prédicteurs ou des critères de niveau ordinal. Les caractéristiques humaines de nature physique (taille, poids, âge, sexe, handicap, etc.) sont presque toujours mesurées soit au niveau nominal, soit au niveau d'intervalle. Quant aux caractéristiques psychologiques (compétences, aptitudes, traits de personnalité, etc.), la plupart sont considérées comme des mesures de niveau d'intervalle (Schneider et Schmitt, 1986), surtout si elles sont mesurées par des tests psychométriques (Dunnette, 1966). De plus, même en présence de prédicteurs ou de critères de niveau ordinal, des

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 3.4
ENSEMBLE DES COMBINAISONS PRÉDICTEUR-CRITÈRE EN FONCTION DE L'ÉCHELLE DE MESURE

		Prédicteur		
		Échelle nominale	Échelle ordinale	Échelle d'intervalle
Critère	Échelle nominale	Combinaison 1	Combinaison 2	Combinaison 3
	Échelle ordinale	Combinaison 4	Combinaison 5	Combinaison 6
	Échelle d'intervalle	Combinaison 7	Combinaison 8	Combinaison 9

auteurs considèrent que la différence entre les résultats obtenus avec l'application de techniques statistiques pour mesures d'intervalle (paramétriques) et ceux obtenus avec les techniques pour mesures ordinales (non paramétriques) est habituellement très faible et ne justifie pas, du moins en pratique, le recours à ces dernières techniques (Baker, Hardick et Petrinovich, 1966). Compte tenu de la rareté des situations mettant en jeu des mesures de niveau ordinal ou considérées comme telles, les combinaisons correspondant à ces situations (2, 4, 5, 6 et 8) ne seront pas présentées.

Une démarche en trois phases. La démarche proposée pour étudier la relation prédicteur-critère est, quelle que soit la combinaison analysée, une démarche en trois phases. La première phase est une étape préliminaire de **vérification** de la qualité des données recueillies pour faire l'étude. La deuxième consiste à examiner directement les données afin de **visualiser** la présence d'une relation et de la **quantifier**, si elle existe, à l'aide de divers indices statistiques. La troisième phase a pour but d'**apprécier la grandeur** de la relation observée en considérant tour à tour des points de vue forts différents.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

PRÉDICTEUR ET CRITÈRE DE NIVEAU D'INTERVALLE (COMBINAISON 9)

Voyons la première situation où le prédicteur et le critère sont tous deux des variables de niveau d'intervalle (combinaison 9). Par exemple, le prédicteur est un nombre d'années de scolarité, un score à un test psychométrique ou un résultat à un examen de connaissances alors que le critère est un volume de ventes, l'évaluation du supérieur sur une échelle en cinq points ou un nombre d'accidents. Comme cette situation est très courante en sélection du personnel, elle servira de prototype à l'étude de la relation prédicteur-critère.

Un exemple réel. Afin d'illustrer le déroulement concret d'une étude de validation, voyons un exemple tiré d'une situation réelle. Une entreprise manufacturière utilise un test d'aptitude mentale générale parmi ses outils de sélection pour l'embauche de ses employés de production. Depuis près de 20 ans, elle fait passer ce test comme exigence de base à ceux qui postulent un emploi. Les candidats dont les scores ne sont pas jugés satisfaisants sont rejetés, alors que les autres sont retenus pour l'étape suivante de sélection. L'entreprise accumule tous les scores des employés qui ont été embauchés depuis le début.

Aujourd'hui, l'entreprise est intéressée à connaître la valeur de ce test d'aptitude comme instrument de sélection. Une étude de validation a été entreprise auprès des 120 employés actuellement affectés à la production ; le rendement de chacun d'eux a été évalué en guise de critère. Cette évaluation du rendement consistait en une appréciation par les supérieurs au moyen d'une grille comportant 10 dimensions (jugement, compréhension du travail, organisation, compétences techniques, etc.). Chaque dimension a été évaluée à l'aide de l'échelle en cinq points suivante : 1) « Insatisfaisant », 2) « Passable », 3) « Satisfaisant », 4) « Supérieur » et 5) «Exceptionnel ». Pour chaque employé, la moyenne obtenue à l'ensemble de ces 10 dimensions constituait son évaluation du rendement. Dans le but d'assurer l'objectivité du critère, l'évaluation du rendement a été faite par deux évaluateurs différents, en présence d'un représentant du personnel. De plus, les résultats demeuraient confidentiels et étaient utilisés au seul bénéfice de la présente étude. L'entreprise se demande

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

s'il y a une relation entre les scores à ce test d'aptitude et le rendement au travail de ses employés de production. Est-ce que ce test est suffisamment valide pour justifier le maintien de son utilisation ?

PHASE 1 VÉRIFICATIONS PRÉLIMINAIRES

Avant d'entreprendre tout travail statistique concernant la relation entre les mesures du prédicteur et celles du critère, il faut vérifier si les données obtenues satisfont à certaines exigences de qualité. L'étude de validation pourrait à la limite être arrêtée, faute de données adéquates. Au moins trois vérifications s'imposent : l'irrégularité dans les données, leur fidélité et leur taux de variation.

Irrégularités dans les données. La première vérification consiste à examiner les données en quête d'irrégularités qui pourraient trahir la présence d'erreurs, de contamination ou même de falsification ; l'expérience nous apprend que l'on n'est jamais trop prudent. Ces irrégularités peuvent provenir des diverses étapes de l'étude : instruments et méthodes de mesure, échantillon, collecte de données, compilation, etc. Les irrégularités devront être corrigées, si possible avant de procéder à l'étude de la relation prédicteur-critère. À défaut de les corriger, l'analyste devra au pire mettre un terme à son étude, au mieux tenir compte de leur présence lors de l'interprétation des résultats.

Une irrégularité est en soi un écart entre ce qui est observé et un état normal, un état souhaité. Repérer des irrégularités présuppose que l'on ait une idée assez claire de la situation dite normale ou souhaitée. Or, à quoi devraient ressembler les mesures obtenues pour le prédicteur et pour le critère ? Pour répondre à cette question, il faut bien connaître la variable mesurée et l'instrument qui a servi à la mesurer. Il faut aussi disposer d'une expérience suffisante avec cet instrument dans le même contexte que celui de l'étude de validation en cours. Présenter les différentes caractéristiques des données pour chaque type d'instruments de mesure des prédicteurs et des critères dépasserait le cadre de cet ouvrage ; revenons plutôt à l'exemple donné plus haut.

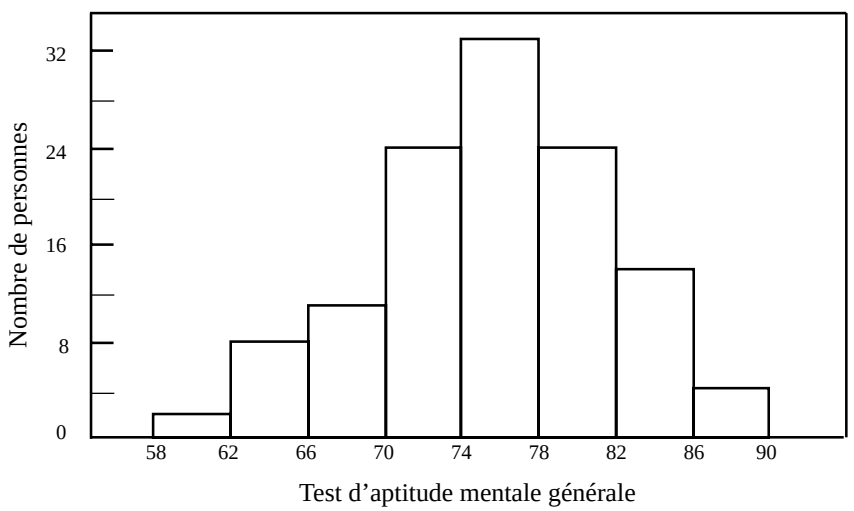
Dans les mesures de niveau d'intervalle, les principaux outils statistiques habituellement utilisés pour vérifier la présence d'irrégularités dans les données sont l'histogramme, la moyenne et l'écart type. L'histogramme des résultats au test d'aptitude mentale générale, soit le **prédicteur**, est reproduit à la figure 3.5. Est-ce que ces données

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

semblent correctes ? Y a-t-il lieu de penser que ces données sont biaisées ou comportent des irrégularités ? L'histogramme indique que les résultats au test d'aptitude pour les 120 travailleurs ont tendance à se distribuer selon une courbe normale. Cette situation n'a rien de surprenant, étant donné que la distribution des scores à ce genre de tests devrait a priori tendre vers une courbe normale lorsqu'un tel test est appliqué à un échantillon représentatif de la population et lorsqu'il n'est ni trop difficile ni trop facile.

Figure 3.5
RÉSULTATS DE 120 EMPLOYÉS D'UNE USINE À UN TEST D'APTITUDE MENTALE GÉNÉRALE



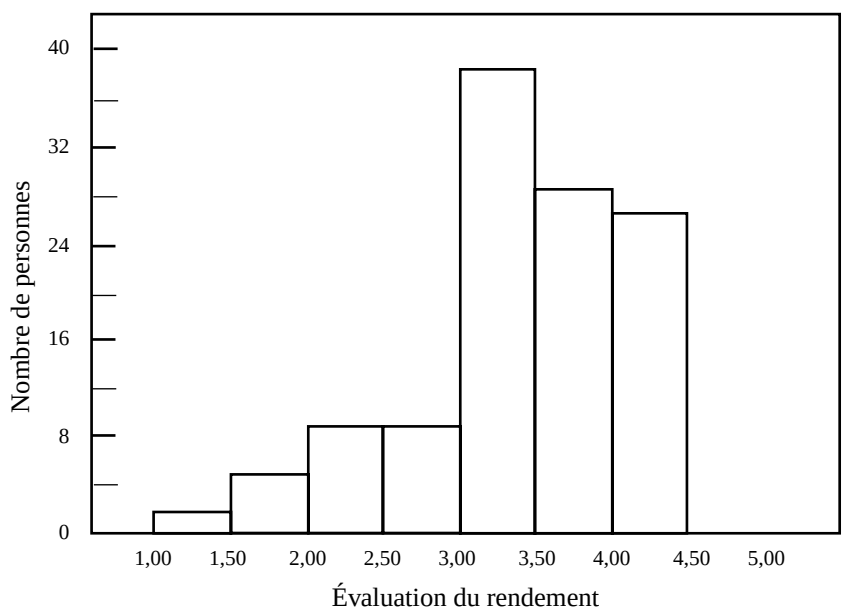
Par ailleurs, nous savons que l'échantillon de la présente étude n'est pas parfaitement représentatif de la population. Il est constitué d'employés en place déjà sélectionnés en raison de leur score à ce test d'aptitude ; conséquemment, les candidats les plus faibles ont déjà été éliminés. Il est donc possible que la moyenne des résultats au test d'aptitude pour ce groupe de travailleurs (moyenne : 74,83) soit légèrement plus élevée que celle de la population générale et que la dispersion (écart type : 6,03) soit plus faible. Après avoir consulté le manuel technique du test concernant les normes de la population générale, les deux hypothèses se sont confirmées. La moyenne de l'échantillon dépasse d'environ 6 points celle de la population générale

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

(qui est d'environ 80) alors que l'écart type est à peu près de 9 points en dessous de celui de la population générale (qui est d'environ 15). La note de passage au test était si élevée qu'il n'est pas nécessairement anormal d'observer une telle différence dans l'écart type chez cet échantillon comparativement à la population générale. En conclusion, les données recueillies auprès de ces 120 travailleurs semblent correctes. La forme de leur distribution, leur moyenne et leur écart type correspondent à ce à quoi l'on pouvait s'attendre en pareille situation. Aucune irrégularité n'est soupçonnée pour le moment.

L'histogramme des évaluations du rendement, soit le **critère**, apparaît à la figure 3.6. La moyenne et l'écart type donnent des valeurs de 3,24 et de 0,72 respectivement. Y a-t-il lieu de penser qu'il y a des irrégularités dans ces données ? L'histogramme indique que les évaluations du rendement pour les 120 travailleurs de l'usine n'ont pas tendance à se distribuer selon une courbe normale : la distribution est fortement asymétrique. Il y a plus de travailleurs qui

Figure 3.6
ÉVALUATION DU RENDEMENT DE 120 EMPLOYÉS D'UNE USINE



© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

obtiennent des évaluations supérieures à la moyenne (M : 3,24) qu'il n'y en a en dessous. Aussi, on remarque qu'aucun travailleur n'a reçu d'évaluation supérieure à 4,50, alors que seulement deux d'entre eux ont plongé sous la barre de 1,50. Que penser de ces données ? Sont-elles correctes ? Qu'est-ce qui pourrait bien expliquer leur déviation par rapport à la loi normale ?

Premièrement, il faut se rappeler que la **distribution normale** ne s'applique pas automatiquement à une évaluation du rendement (voir chapitre 7, section « Déviation par rapport à la distribution normale »). Deuxièmement, cet échantillon est loin d'être **représentatif** de la population générale ; les travailleurs retenus dans cet échantillon ont été sélectionnés pour leurs compétences lors de leur embauche. De plus, ces personnes ont, à des degrés variables selon leur expérience, eu la chance d'améliorer ces compétences. Enfin, cet échantillon avait ceci de particulier : il ne comprenait que la moitié des employés de production de l'usine, soit ceux qui avaient été promus aux postes les plus élevés et les mieux rémunérés. En conclusion, il faudrait plus d'informations pour se faire une idée de la distribution attendue ou souhaitée et se prononcer sur la présence d'irrégularités. Comme on peut le constater, la réponse est plus une affaire de connaissances et de jugement que de chiffres et de statistiques. Il est rare qu'une calculatrice fournisse des réponses à une question le moindrement complexe.

Fidélité du prédicteur et du critère. La deuxième vérification porte sur la fidélité des données. Dans toute mesure, des erreurs purement accidentelles risquent de se produire : inattention du répondant, ambiguïté dans une question ou dans des directives, fatigue du correcteur ou de l'évaluateur, bruit soudain, etc. Moins il y a d'erreur de ce type, plus les données sont fidèles. Or, la présence de ces erreurs, au critère comme au prédicteur, diminue la relation entre ces deux variables. La validité d'un instrument de mesure est ainsi limitée par sa propre fidélité et celle du critère ; la fidélité constitue par le fait même la valeur maximale pouvant être atteinte pour le coefficient de validité (voir chapitre 4).

Connaître ainsi le plafond théorique de la validité fournira un point de référence additionnel permettant d'apprécier la grandeur du coefficient de validité qui sera estimée à l'étape suivante. Cet

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

aspect sera repris plus en détail ultérieurement (voir section « Appréciation de la grandeur du coefficient de validité »). Pour l'instant, retenons que la fidélité du test d'aptitude mentale générale, utilisé dans l'exemple des 120 employés de production en usine, a été estimée à 0,91 à l'aide d'une méthode de test-retest (Kellogg et Norton, 1978). Quant à l'évaluation du rendement, un coefficient de fidélité de 0,83 a été obtenu par la méthode de consistance interne alpha. Ces indices de fidélité n'imposeront que peu de limitation au coefficient de validité estimé à l'étape suivante. L'étude de validation peut se poursuivre, parce que les mesures sont suffisamment fidèles pour permettre l'existence d'une relation entre le prédicteur et le critère. Nous verrons plus loin qu'il existe des formules statistiques pour compenser le manque de fidélité.

Variation ou restriction de l'étendue. La troisième vérification vise à assurer qu'il y a suffisamment de variation dans les résultats de chacune des deux variables, prédicteur et critère. Autrement dit, chacune des distributions met-elle en évidence des différences entre les individus ? S'il n'y a pas suffisamment de variation dans les résultats de l'une ou l'autre des variables, il est inutile d'essayer de déceler une relation entre un instrument de sélection et un critère de rendement. Avec des individus sensiblement égaux sur l'une ou l'autre des variables, la corrélation tendrait vers zéro. Par exemple, il est impossible de vérifier la relation entre le sexe des étudiants (prédicteur) et le rendement scolaire (critère), si presque tous les étudiants de l'échantillon sont de sexe féminin. De même, la relation entre l'expérience des travailleurs (prédicteur) et le rendement au travail (critère) ne peut être étudiée si les travailleurs de l'échantillon ont tous un rendement jugé satisfaisant.

Il y a parfois des situations où la variation au prédicteur ou au critère a été particulièrement réduite comme dans les cas suivants : 1) les résultats à l'instrument de sélection à l'étude ont servi pour éliminer les candidats les moins aptes, 2) le schème concomitant repose sur un échantillon d'employés dont une partie des plus faibles et des plus forts ont été retranchés, 3) la mesure du critère repose sur l'évaluation du supérieur et elle est sujette à une procédure de règlement des griefs ou 4) les meilleurs employés ont été promus ou ont quitté pour de meilleures conditions dans d'autres organisations.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Mais comment savoir s'il y a suffisamment de variation dans les mesures du prédicteur et du critère ? Une façon d'y arriver est de comparer la variation obtenue à chacune des variables (prédicteur et critère) à la variation théorique qui devrait normalement être obtenue dans une population comparable à l'échantillon. Si l'on connaît cette variation théorique, on n'a qu'à appliquer une formule d'atténuation pour estimer la corrélation réelle. En effet, comme pour le manque de fidélité des mesures, des formules permettent de compenser l'influence de la baisse de variation sur la grandeur du coefficient de validité obtenu. Le manque de variation et ses effets sur la validité sont décrits dans la documentation sous le terme de **restriction de l'étendue** de l'échantillon. Cet aspect est vu en détail plus loin dans ce chapitre (voir section « Appréciation de la grandeur du coefficient de validité »).

Revenons maintenant à notre exemple. Au regard de la variation du **prédicteur**, on peut apprendre, en consultant le manuel technique accompagnant ce test, que l'écart type de ce test appliqué à une population générale est d'environ 15 points et que les scores varient habituellement de 6 à 92. Comparativement à la population générale, il y a donc nettement moins de variation dans l'échantillon des 120 travailleurs à l'étude, dont l'écart type est de 6,03 et les scores varient de 58 à 90. La variation existe et justifie la poursuite de l'étude de validité. Il faudra néanmoins tenir compte de cette restriction de l'étendue du prédicteur lors du calcul du coefficient de validité en appliquant la formule d'atténuation appropriée.

Au regard du **critère**, la restriction de l'étendue de la variation est plus difficile à estimer, car on ne connaît pas la variation théorique ou attendue pour cette population d'employés où il n'y aurait eu aucune restriction, par exemple en raison du roulement du personnel ou des promotions ; nous devons tout de même vérifier s'il y a suffisamment de variation dans les mesures, parce que sans cette variation, il est impossible d'observer de relation entre le prédicteur et le critère. L'examen des résultats à l'évaluation du rendement des 120 travailleurs de l'échantillon révèle que ces scores varient de 1,0 à 4,5, et se dispersent sur presque toute l'étendue de l'échelle de 1,0 à 5,0 mise à la disposition des évaluateurs. L'écart type est de 0,72, ce qui donne un coefficient de variation de 22,2 %, soit l'écart type (0,72) divisé par la moyenne (3,24) et multiplié par 100. Il y a donc de la variation

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

au critère. Ce résultat se compare à la variation moyenne de rendement estimée par Hunter, Schmidt et Judiesch (1990) autour de 19 % dans ce genre d'emplois peu complexes (ouvriers non spécialisés ou semi-spécialisés).

PHASE 2 DIAGRAMME DE DISPERSION ET COEFFICIENT DE VALIDITÉ

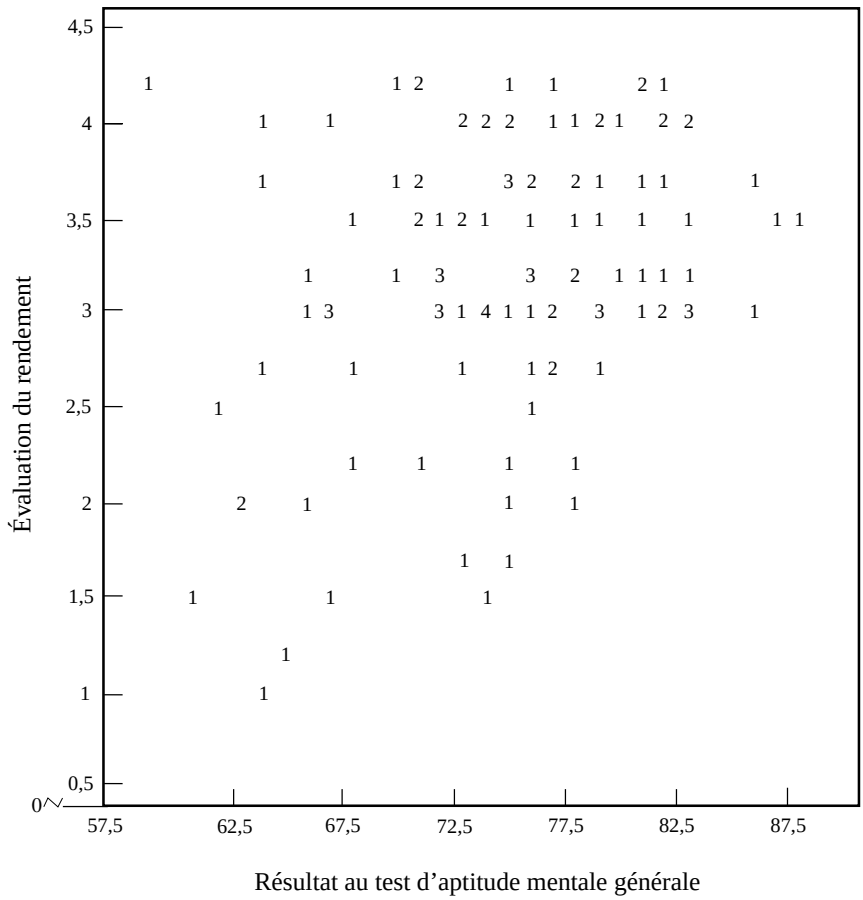
Les vérifications préliminaires ont permis de s'assurer que les mesures du prédicteur et du critère semblent adéquates : 1) aucune irrégularité soupçonnée, 2) des coefficients de fidélité acceptables et 3) de la variation dans les mesures de chacune des variables. Ces résultats autorisent la poursuite de l'analyse pour voir s'il y a une relation entre le prédicteur et le critère de rendement.

Diagramme de dispersion. Toute analyse de validité critériée, lorsque les données sont de niveau d'intervalle, doit débuter par la compilation d'un diagramme de dispersion ; c'est un moyen à la fois simple et efficace de visualiser directement la relation entre deux ensembles de données (voir chapitre 7, section « Diagramme de dispersion »). À partir des données de l'exemple précédent concernant les 120 employés de l'usine, le diagramme de dispersion entre leurs scores au test d'aptitude mentale générale et leur évaluation du rendement a été compilé (voir figure 3.7). L'examen du diagramme laisse entrevoir une répartition des points non aléatoire. Au contraire, la relation entre le prédicteur et le critère a tendance à être de forme linéaire : plus les résultats au test d'aptitude sont élevés, plus le rendement est élevé. En effet, on constate que parmi les employés dont le score au test était très faible, par exemple inférieur à 70,0, plusieurs ont obtenu une évaluation de leur rendement inférieure à 3,0. Cette proportion diminue de façon notable à mesure que les employés détiennent un score plus élevé au test, de sorte que les employés dont le score au test est supérieur à la moyenne (soit un score d'environ 75,0 et plus) ont presque tous obtenu une évaluation de leur rendement de 3,0 et plus. Étant donné cette relation, ce test d'aptitude mentale pourrait permettre à l'organisation d'améliorer le rendement des nouveaux employés de production en sélectionnant ceux dont le score au test est le plus élevé.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 3.7
DIAGRAMME DE DISPERSION ENTRE UN TEST D'APTITUDE MENTALE GÉNÉRALE ET L'ÉVALUATION DU RENDEMENT POUR 120 EMPLOYÉS D'UNE USINE



Note : Dans le diagramme, les chiffres indiquent le nombre de personnes ayant exactement les mêmes coordonnées.

Coefficient de corrélation. Comme la relation prédicteur-critère exposée à la figure 3.7 a tendance à être linéaire, la grandeur de cette relation peut être quantifiée en compilant le coefficient de corrélation. La valeur obtenue est de 0,31 ; cet indice est le coefficient de validité critériée (prédictive) de ce test d'aptitude mentale générale.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

PHASE 3 APPRÉCIATION DE LA GRANDEUR DU COEFFICIENT DE VALIDITÉ

Comment juger de la grandeur du coefficient de validité ? Disons qu'il y a plusieurs façons de répondre à cette question. Chacune d'elles aborde la question d'un point de vue différent et permet d'accumuler une foule d'informations. Ensemble, ces informations permettront une appréciation nuancée et réaliste de la grandeur de la validité.

A) Recourir au cadre statistique

La grandeur d'un coefficient de validité peut être appréciée en des termes plus ou moins absolus, suivant la logique de la théorie statistique (voir chapitre 7, section « Coefficients de corrélation et de détermination »). Par exemple, on peut situer ce coefficient de 0,31 sur l'**échelle de 0 à 1,0** définissant l'étendue théorique des coefficients de corrélation. Ou bien, il suffit d'élever la corrélation au carré pour obtenir le **coefficient de détermination**. On apprend alors qu'environ 9,6 % de l'évaluation du rendement est reliée à l'aptitude mentale telle qu'elle est mesurée par le test. Enfin, il est possible d'avoir recours à un **test statistique**. Ainsi, la table statistique de signification (voir appendice A, tableau A.2) indique que la probabilité d'obtenir par hasard une corrélation de 0,31 (ou plus) est inférieure à 1 chance sur 100. En effet, pour un échantillon de 100 sujets, la table indique une valeur de 0,25 dans 1 % des cas. En conséquence, il y a moins de 1 chance sur 100 de se tromper lorsqu'on affirme qu'il y a réellement une relation entre les scores à ce test d'aptitude et l'évaluation du rendement chez ces employés. Toutefois, il faut signaler que le fait que la corrélation ne soit pas due au hasard ne change pas un aspect important de la réalité : ce test ne prédit que 9,6 % du rendement chez l'échantillon étudié, ce qui laisse plus de 90 % du critère de rendement non expliqué. Effectivement, en valeur absolue, la grandeur du coefficient de validité peut sembler décevante.

B) Consulter les autres études locales de validation

Un instrument de sélection ne peut pas être évalué par rapport à un standard de perfection, mais plutôt par rapport à ce qu'il peut réaliser en pratique (*Standards*, 1999). Voilà un autre aspect à considérer avant de se faire une idée concernant la grandeur de la validité. Il s'agit de comparer le coefficient de validité de l'instrument avec ceux obtenus

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

dans d'autres études similaires ou avec ceux obtenus par d'autres instruments qui auraient pu être utilisés dans cette situation. Il s'agit alors de consulter la documentation spécialisée, de repérer ces études et de prendre connaissance des coefficients de validité rapportés. Cette approche permettra non seulement de relativiser la grandeur d'un coefficient donné, mais aussi de jauger ce qui est réaliste d'espérer dans un contexte particulier.

Il peut être onéreux de faire une recension bibliographique exhaustive d'autres études de validation correspondant au contexte particulier de la situation. Toutefois, il existe déjà des revues de littérature qui résument un grand nombre de ces études empiriques. Il s'agit donc de consulter les principales revues pour avoir une idée assez claire des coefficients de validité moyens déjà obtenus. La majorité de ces revues fournissent les résultats par types d'instrument de sélection ; certaines d'entre elles tiennent compte de paramètres importants comme la nature des emplois, le schème utilisé pour la collecte des données, le type de critère de rendement, etc. Ne pouvant en faire la liste complète, nous allons au moins mentionner la plus importante.

Les compilations de Ghiselli. La revue la plus exhaustive d'études de validation concernant les tests psychométriques est certainement celle de Ghiselli **(1966)**. Cet auteur s'est attelé à la tâche fastidieuse de recenser à peu près toutes les études réalisées entre 1920 et 1964 portant sur les tests, pour ensuite compiler les coefficients de validité moyens. Les résultats sont présentés sous forme de tableau par types de tests et par catégories d'emplois. Par exemple, Ghiselli rapporte des coefficients de validité moyens de près de 0,25 pour les tests d'aptitudes mentales, perceptuelles ainsi que pour les tests de personnalité utilisés pour la sélection du personnel cadre. Pour le personnel de supervision, il obtient des coefficients similaires avec les tests d'aptitudes cognitives, spatiales et mécaniques. Pour les postes de bas niveau dans la vente (p. ex., commis vendeurs), les tests d'aptitudes cognitives n'ont que peu de validité. En revanche, les tests de personnalité obtiennent des coefficients moyens d'environ 0,35 avec un critère de rendement au travail. Pour les postes de vendeurs de haut niveau (p. ex., représentant), la validité des tests de personnalité baisse autour de 0,27, alors que pour les tests d'aptitudes mentales la validité est d'environ 0,31.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Ghiselli mentionne que, pour l'ensemble des postes, la validité moyenne des tests à prédire le succès lors de la formation en emploi est de 0,30, alors qu'elle passe à 0,19 lorsqu'il s'agit de prédire le rendement effectif en emploi. Cette différence de 0,11 en faveur d'un critère d'apprentissage est observée pour presque tous les postes, mais pas pour tous les types de tests. Les tests d'aptitudes cognitives, spatiales et mécaniques prédisent mieux le succès en formation que le rendement effectif, alors que les tests d'aptitudes perceptuelles et motrices ont des validités d'égale valeur pour les deux types de critère[11]. Ghiselli, en **1973**, a publié une seconde revue semblable à la première, quoique de moindre envergure. Cette fois, la validité moyenne a été de 0,45 par rapport au critère de succès en formation et de 0,35 par rapport au rendement effectif. Ces quelques résultats au sujet des travaux de Ghiselli (1966, 1973) sont tirés d'un résumé présenté par Wigdor et Garner (1982).

Revenons maintenant à **l'exemple** des 120 employés où la corrélation de 0,31 a été obtenue entre un test d'aptitude mentale générale et un critère de rendement effectif en emploi. Alors, comment se compare ce coefficient de validité avec les autres études en des circonstances similaires ? Ghiselli (1973) obtient des corrélations moyennes autour de 0,20 entre un critère de rendement effectif (comme dans l'exemple) et les tests d'aptitudes cognitives appliqués à des travailleurs de l'industrie (échantillon de plus de 10 000 sujets). Les tests d'aptitudes mécaniques et spatiales donnent à peu près les mêmes résultats. Une autre étude, plus générale, réalisée par Coward et Sackett (1990) sur 174 échantillons représentant une grande variété de postes et totalisant 36 614 sujets, rapporte des corrélations moyennes de 0,21 pour les tests d'aptitudes. Les corrélations sont de 0,16 pour l'aptitude verbale, de 0,21 pour l'aptitude numérique et de 0,14 pour les aptitudes spatiales et perceptuelles. À la lumière de ces résultats, la réponse à notre question serait que le coefficient de 0,31 obtenu dans notre exemple est un peu supérieur à la moyenne.

---

11. Une étude de Ree et Earles (1991), menée auprès de 78 041 militaires répartis dans 82 types d'emplois, confirme cette tendance : la prédiction du succès lors de la formation est surtout affaire d'aptitude mentale générale (ou facteur *g*, particulièrement présent dans les tests d'aptitudes mentales et spatiales), alors que les aptitudes spécifiques n'y ajoutent pas grand-chose.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

C) Considérer les faiblesses méthodologiques

D'après ces résultats, il semble que la validité des tests psychométriques à prédire le rendement au travail ne soit pas très élevée. En fait, l'ensemble des études menées dans les années 1950 et 1960 laissaient croire que les coefficients de validité même pour les meilleurs outils de sélection avaient tendance à plafonner médiocrement aux alentours de 0,30 (Smith, Gregg et Andrews, 1989). Or, nous savons aujourd'hui que de nombreuses études locales de validation recensées à l'époque de Ghiselli présentent des faiblesses méthodologiques qui ont eu pour effet de sous-estimer substantiellement les coefficients de validité. Il est donc important d'examiner la méthodologie suivie lors d'une étude de validation pour voir dans quelle mesure les coefficients de validité sont évalués à leur pleine valeur.

1. Restriction de l'étendue. La première faiblesse qui peut diminuer la valeur du coefficient obtenu est ce qu'il est convenu d'appeler la restriction de l'étendue de l'échantillon. Pour observer une corrélation entre deux variables, nous savons qu'il doit y avoir de la variation dans les mesures obtenues pour le prédicteur et pour le critère. En d'autres mots, les personnes doivent être différentes les unes des autres par rapport à ces deux variables. S'il n'y a pas suffisamment de variation, il ne sera pas possible de voir de relation entre un instrument de sélection et un critère de rendement. Par exemple, il est impossible de vérifier la relation entre l'ancienneté à une occupation (prédicteur) et le rendement (critère) si les personnes étudiées ont toutes le même nombre d'années d'ancienneté. Le peu de variation entre les résultats obtenus pour les variables étudiées affectera donc à la baisse la valeur du coefficient de corrélation.

Il est facile de comprendre ce phénomène à partir d'un exemple. Prenons une étude de validation suivant un schème concomitant où les résultats au test à valider auraient servi lors de la sélection. En éliminant ainsi les candidats les plus faibles dès l'embauche, on détruit une partie de la preuve qui aurait vraisemblablement servi à démontrer que des scores faibles aux tests sont reliés à un rendement effectivement plus faible. Pour illustrer davantage, supposons que le directeur du module en sciences comptables ait pris connaissance de la relation entre la moyenne cumulative universitaire et les résultats

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

aux examens de l'Ordre des comptables agréés (voir figure 7.11) et qu'il décide de mettre cette information à profit. Lorsque les nouveaux finissants viennent lui demander conseil pour savoir s'ils devraient ou non s'inscrire aux examens de l'Ordre, le directeur du module décourage très fortement les étudiants ayant une moyenne cumulative inférieure à 3,0 à s'inscrire aux examens, en leur disant par exemple que leurs chances de succès ne valent pas les innombrables efforts à consentir ou qu'un échec aux examens pourrait compromettre l'obtention d'un emploi. Le directeur a été si convaincant que la majorité des finissants visés ne se sont pas inscrits. À l'aide du diagramme de dispersion (voir figure 7.11), et en supposant que ces résultats sont représentatifs de la situation actuelle, voyons quel genre de relation il y aurait s'il ne restait que les finissants dont la moyenne universitaire est de 3,0 et plus. Le nuage de points restants ne formerait pas aussi nettement une ellipse, caractéristique d'une relation très forte. D'ailleurs, la corrélation pour ces 14 finissants ne serait que de 0,28 comparativement à 0,77 qu'elle était avec les 46 finissants.

En sélection, une restriction de l'étendue de l'échantillon arrive fréquemment. Elle peut survenir au niveau du prédicteur, au niveau du critère ou aux deux à la fois (Arvey et Faley, 1988). La restriction de la variation du **prédicteur** se produit, par exemple, lorsque les résultats à l'instrument de sélection à l'étude ont servi lors de l'embauche pour éliminer les candidats les moins aptes. La restriction peut aussi survenir lorsque les candidats ont été éliminés à l'embauche à partir d'exigences reliées au prédicteur à l'étude. Ainsi, dans la majorité des emplois techniques, professionnels et de gestion, le rejet des candidats non performants a déjà été effectué par le biais de la réussite ou de l'échec aux examens des établissements d'enseignement conduisant à l'obtention du diplôme requis. Par conséquent, une étude de validation portant sur un test d'aptitude mentale générale serait vraisemblablement affectée par la restriction des scores des candidats à ce test (Barrette et Durivage, 1993).

Quant à la restriction au niveau du **critère**, elle peut être amenée par le roulement du personnel, les promotions ou les affectations qui ont eu lieu avant la collecte des données, comme c'est souvent le cas pour un schème concomitant de validation avec les employés en

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

place. Par exemple, si d'une part les meilleurs employés ont été promus ou ont quitté pour bénéficier de meilleures conditions dans d'autres organisations et que d'autre part les employés qui ne présentaient pas un rendement adéquat ont été amenés à quitter ou sont partis de leur propre gré, il y aura une diminution de la variation au rendement. Il arrive également que l'on observe de la restriction au critère lorsque sa mesure repose sur une évaluation du supérieur, évaluation soumise à des pressions de toutes sortes comme une procédure d'appel ou de règlement des griefs.

S'il y a restriction de l'étendue de l'échantillon comparativement à la population générale des candidats, les coefficients de validité observés sont plus faibles. En pareille situation, il faut ajuster les coefficients de validité dans la mesure du possible (SIOP, 1987). Il existe des formules qui permettent de compenser ces restrictions et de connaître ainsi la corrélation qui aurait été obtenue si aucune restriction n'avait affecté l'échantillon. Par exemple, voici la formule qui estime la validité réelle lorsqu'on désire éliminer l'effet de la restriction au **prédicteur** (Thorndike, 1949, cité dans Guion, 1965, 1998).

Validité si prédicteur non restreint $\left(r_{x_{nr}y}\right) = \dfrac{r_{xy}\left(S_{x_{nr}} / S_x\right)}{\sqrt{1 - r_{xy}^{\,2} + r_{xy}^{\,2}\left(S_{x_{nr}}^{\,2} / S_x^{\,2}\right)}}$

où r_{xy} : validité observée après restriction au prédicteur

$S_{x_{nr}}$: écart type du prédicteur (x) non restreint ou attendu

S_x : écart type observé après restriction au prédicteur

Ce genre de faiblesse est donc facile à contourner, mais à condition de connaître la restriction au prédicteur et au critère. Cette information n'est malheureusement pas toujours disponible, ce qui peut expliquer que la correction pour restriction de l'étendue de l'échantillon est très peu appliquée en pratique (Ree *et al.*, 1994). Pour le prédicteur, s'il s'agit de tests connus ou édités, on peut connaître l'étendue non restreinte des résultats en consultant les normes publiées. En effet, l'écart type de la population générale fournit une

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

estimation tout à fait acceptable (Hoffman, 1995). Dans l'exemple des employés d'usine, le test d'aptitude mentale générale avait servi à sélectionner des candidats lors de l'embauche. Par conséquent, il y a une restriction au niveau du prédicteur, qui a été estimée à environ 9 points d'écart type : l'écart type de l'échantillon est de 6,07 alors que celui de la population générale est de 15 (voir section « Vérifications préliminaires »). Sans cette restriction au prédicteur, la validité observée n'aurait pas été de 0,31, mais vraisemblablement de 0,63, ce qui est considérablement plus élevé. Les calculs sont présentés ci-dessous :

$$\frac{0{,}31(15/6{,}03)}{\sqrt{1-0{,}31^2+0{,}31^2\left(15^2/6{,}03^2\right)}}=\frac{0{,}7711}{1{,}2242}=62{,}99$$

Il y a probablement aussi une restriction au niveau du **critère** de rendement, étant donné que l'ancienneté moyenne des employés de l'échantillon est de plus de 6 ans ; plusieurs d'entre eux totalisent même près de 20 ans de service. Au cours d'une si longue période, il y a lieu de croire que de nombreux employés ne sont plus à leur emploi parce que leur profil ne correspondait pas exactement aux exigences, créant du coup un échantillon plus homogène par rapport au rendement. Sans avoir été estimée directement, la restriction au niveau des évaluations du rendement de ces 120 employés tend à être confirmée par la distribution fortement asymétrique observée plus tôt (voir section « Vérifications préliminaires »). Mais il est impossible d'estimer précisément cette restriction, parce que l'écart type de la population de ces employés pour le critère de rendement n'est pas connu.

2. Fidélité du prédicteur et du critère. La deuxième faiblesse méthodologique qui peut fausser à la baisse l'estimation de la validité provient de la fidélité des mesures. En effet, rappelons qu'un prédicteur ne peut pas corréler avec une autre mesure plus fortement qu'avec lui-même (McCormick et Tiffin, 1974). Prenons l'exemple d'un prédicteur dont la fidélité est de 0,80. La prédiction la plus forte qu'il est possible d'obtenir entre ce prédicteur et un critère est donc de 80 % de variance expliquée, ce qui veut dire que la corrélation maximale observée serait de 0,89 (soit la racine carrée de 0,80) (voir chapitre 4, section « Plafond sur la validité et correction pour atténuation »). De la même façon, un prédicteur ne peut pas prédire un critère de rendement mieux que le critère lui-même. Bref, la validité d'un

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

instrument de mesure est limitée par sa fidélité et celle du critère. La relation entre la validité et la fidélité est clairement exprimée par la formule suivante (Guion, 1998) :

Validité si critère et prédicteur parfaitement fidèles $\left(r_{x\infty\, y\infty}\right) = \dfrac{r_{xy}}{\sqrt{r_{xx}\, r_{yy}}}$

où r_{xy} : validité observée entre le prédicteur et le critère

r_{xx} : fidélité observée du prédicteur (x)

r_{yy} : fidélité observée du critère (y)

Grâce à cette formule, il est aisé de calculer la validité théorique ($r_{x\infty y\infty}$) qui existerait si le prédicteur (x) et le critère (y) étaient parfaitement fidèles, ce qui permet de corriger en quelque sorte la validité obtenue pour l'infidélité du prédicteur et du critère. Appliquons cette formule à l'exemple des 120 travailleurs d'usine. Sachant que la fidélité est de 0,91 au prédicteur et de 0,83 au critère, la validité obtenue de 0,31 devient 0,36, soit la validité que l'on aurait dû obtenir si les mesures du prédicteur et du critère avaient été parfaitement fidèles. C'est ce qu'on appelle la correction pour l'atténuation du prédicteur et du critère.

Bien qu'elle soit intéressante d'un point de vue théorique pour estimer la relation réelle entre deux variables, l'application de cette formule est de peu d'utilité en pratique. En effet, Guion (1965, 1998) rappelle qu'il ne sert à rien de rêver, car aucun prédicteur n'est parfaitement fidèle. Cependant, même s'il n'y a pas plus de critères que de prédicteurs parfaitement fidèles, il importe d'estimer la validité qui serait obtenue si le critère de rendement avait été parfaitement fidèle. En effet, calculer la validité du prédicteur seulement par rapport à la portion « explicable » du rendement, en excluant la portion d'erreur aléatoire, fournit une estimation plus exacte de la capacité prédictive réelle d'un instrument de sélection. C'est d'ailleurs la recommandation de la Society for Industrial and Organizational Psychology (1987, p. 16) : ajuster les coefficients de validité pour

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

l'infidélité au regard du critère seulement, si toutefois il existe un tel estimé de la fidélité qui soit approprié[12]. La formule pour l'atténuation du critère seulement est (Guion, 1965, 1998) :

Validité si critère parfaitement fidèle $\left(r_{xy\infty}\right) = \dfrac{r_{xy}}{\sqrt{r_{yy}}}$
où r_{xy} : validité observée entre le prédicteur et le critère r_{yy} : fidélité observée du critère (y)

Appliquée à notre exemple, la validité observée de 0,31 deviendrait 0,34 si la fidélité du critère, actuellement de 0,83, était parfaite.

3. Erreur d'échantillonnage. Il existe une troisième faiblesse méthodologique qui affecte la précision de l'estimation de la validité lors d'une étude de validation : c'est l'erreur d'échantillonnage. Contrairement aux deux premières faiblesses qui avaient pour effet de sous-estimer la validité réelle en diminuant le coefficient observé, l'erreur d'échantillonnage est aléatoire. Par conséquent, le coefficient de validité obtenu est parfois plus grand qu'en réalité, parfois plus petit. C'est l'erreur inévitable qui est due au hasard lorsqu'une étude porte sur un échantillon plutôt que sur la population entière.

Pourtant, les études de validité publiées reposent sur des échantillons qui ont tendance à être de petite taille. Lent *et al.* (1971, cité dans Cook, 1988) rapportent que la taille médiane des échantillons observés dans 406 études de validité est de 68 sujets. Or, une corrélation calculée sur un si petit échantillon est très peu fiable. Supposons par exemple que la corrélation réelle dans l'ensemble de la population est de 0,30 et qu'un échantillon de 68 candidats est tiré au hasard, il y a une chance sur trois que la corrélation obtenue soit de 0,18 et moins ou de 0,42 et plus (Schmidt *et al.*, 1985). Avec une

12. À cet égard, Austin et Villanova (1992) font remarquer que Schmidt, Hunter et leurs collaborateurs préconisent une valeur de 0,60 comme estimation de la fidélité lorsque vient le temps d'effectuer leur correction pour l'infidélité du critère dans les études de validation, comparativement à 0,80 pour les prédicteurs.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

telle marge d'erreur, le danger est très grand de ne pas détecter une validité réelle (erreur de type II) ou d'observer une validité là où il n'y en a pas en réalité (erreur de type I). Il faut donc un échantillon le plus grand possible.

Alors, de quelle taille doit être un échantillon dans une étude de validité ? Sans préciser outre mesure, Guion (1991) affirme que les échantillons doivent certainement être beaucoup plus grands que ceux rapportés dans les études habituelles et qui comportent de 30 à 50 sujets. Nous avons vu que Schmidt, Hunter et Urry (1976) recommandent un échantillon de 172 sujets et plus pour avoir une chance raisonnable de détecter une validité réelle d'ampleur moyenne (corrélation corrigée d'environ 0,50). Smith, Gregg et Andrews (1989) préconisent plus de 200 sujets. D'autres auteurs ont publié au sujet de la taille requise de l'échantillon (Raju, Edwards et Loverde, 1985 ; Sacket et Wade, 1983, cité dans Guion, 1991). Dans notre exemple des employés de l'usine, l'échantillon compte 120 personnes ; la taille de cet échantillon est substantielle, sans être toutefois pleinement satisfaisante : un ajout d'une cinquantaine d'employés aurait été encore mieux.

4. Contamination du critère. Une quatrième faiblesse concerne la contamination du critère. Un critère peut être biaisé par une foule de facteurs : un superviseur peut avoir un parti pris lors de l'évaluation du rendement de ses subordonnés, la valeur mensuelle des ventes d'un représentant peut être affectée par la faillite d'un gros client situé sur son territoire, le nombre de nouveaux abonnés peut avoir été augmenté au détriment de la marge bénéficiaire, etc. En fait, il y a rarement un seul critère qui soit à lui seul exactement représentatif du rendement idéal attendu de l'organisation. Il y a dans la plupart des cas une possibilité de biais et il n'est pas toujours possible de déterminer dans quel sens le critère est biaisé, si c'est en plus ou en moins. Cela dépend du critère, des biais en cause et même de chacun des employés.

Dans une étude de validation, si les biais sont corrélés au prédicteur, la validité obtenue sera artificiellement plus élevée que la réalité, c'est-à-dire surestimée ; inversement, si les biais ne sont pas corrélés au prédicteur, la validité sera sous-estimée. Par exemple, supposons que le prédicteur est le résultat à une entrevue structurée menée conjointement par le superviseur et un représentant du service

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

du personnel, et que le critère est une évaluation du rendement basée sur le jugement du même superviseur au terme d'une période de probation d'un mois. Il est facile d'imaginer que les préjugés du superviseur lors de l'entrevue (p. ex., préférence naturelle pour une personne de telle origine ethnique) auront tendance à être présents lors de l'évaluation du rendement. Le meilleur remède à la contamination du critère est de choisir dès le départ un critère représentatif du rendement souhaité et résistant aux biais. Si un biais était découvert après coup, il est parfois possible d'en diminuer l'influence sur la validité par des corrections statistiques (Guion, 1965)[13].

Dans l'exemple des employés d'usine, le prédicteur est un test d'aptitude administré par le service du personnel et dont les résultats sont tenus confidentiels. Les supérieurs n'ont pas eu accès à cette information afin d'éviter un biais possible lors de l'évaluation du rendement qui servait de critère dans l'étude de validation. Pour contrôler encore davantage l'objectivité du critère, rappelons que l'évaluation du rendement a été réalisée par deux évaluateurs différents, en présence d'un représentant du service des ressources humaines. De

13. En contexte de validation d'instruments de sélection, il faut toujours avoir à l'esprit la possibilité de contamination du critère de rendement, sachant que la nature même du critère peut influencer le coefficient de validité obtenu. Par exemple, des critères objectifs de rendement donnent souvent lieu à des coefficients de validité différents de ceux obtenus à l'aide de critères subjectifs. Ainsi, après analyse de Validity Information Exchange paru dans *Personnel Psychology* pour une période de 20 ans, les coefficients de validité ont été significatifs dans 82 % (205/250) de toutes les études utilisant un critère objectif, contre 40 % (377/948) de celles ayant un critère subjectif et 25 % (47/177) de celles dont le critère subjectif était indirect (promotions, niveau salarial, etc.) [Lent, Aurbach et Lewin, 1971]. Schmitt *et al*. (1984) ont remarqué aussi que les coefficients de validité peuvent varier selon le type de critère utilisé, sans toutefois observer d'effet systématique selon la nature objective ou subjective du critère. Par contre, d'autres études n'ont pas trouvé de différences importantes de validité entre un indice objectif de productivité et une évaluation subjective du rendement effectuée par le supérieur (Hoffman, Natham et Holden, 1991 ; Natham et Alexander, 1988). Dans certains cas, c'est le contraire. Le centre d'évaluation, qui utilise largement la simulation et le jugement d'observateurs, rapporte des coefficients de validité plus élevés lorsque le critère de rendement est un jugement global porté sur le potentiel de gestion du candidat (Gaugler *et al.*, 1987). Peut-être y a-t-il interaction entre la nature du critère et la nature du prédicteur ? Ou peut-être s'agit-il simplement d'une contamination du critère dans le cas du centre d'évaluation (Dreher et Sackett, 1983) ? D'autres études seraient nécessaires.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

plus, les résultats de l'évaluation du rendement étaient gardés confidentiels et utilisés au seul bénéfice de la présente étude, donc ils ne pouvaient en aucune façon être consultés pour la gestion des employés.

5. Traitement statistique. Finalement, le traitement statistique doit être approprié à la nature des données. D'abord, chaque test statistique est basé sur des postulats concernant les données recueillies. Par exemple, le calcul d'une simple corrélation produit-moment de Pearson, qui est la statistique de loin la plus utilisée dans les études de validation, exige que la relation entre le prédicteur et le critère soit linéaire. Si cette condition n'est pas respectée, la corrélation obtenue témoignera d'une relation moins élevée que la réalité. Il faut donc vérifier ce postulat de linéarité au minimum en traçant le diagramme de dispersion[14].

Il arrive aussi que des données de nature quantitative (p. ex., les notes à un examen de connaissances) soient ramenées sur une échelle ne comportant que quelques points (p. ex., « succès-échec » ou « fort-moyen-faible ». Il faut savoir que cette pratique, peu recommandable, fait perdre de l'information, réduit substantiellement l'estimation de la variance expliquée et, conséquemment, diminue la puissance des tests statistiques (Guion, 1998). Prenons le cas d'une variable dont les données seraient scindées de part et d'autre de la moyenne, réduisant ainsi les valeurs obtenues à deux possibilités (soit « en haut » et « en bas » de la moyenne). Si une corrélation est calculée avec cette nouvelle variable dichotomisée, un tel coefficient sera réduit par un facteur de 0,798, soit environ 80 % (Magnusson, 1966). Ainsi, une corrélation de 0,30 entre deux variables continues deviendrait 0,24 (ou 0,30 multiplié par 80 %) si les données d'une de ces variables sont ramenées à deux valeurs. Selon Guion (1991), l'utilisation d'une note de passage peut justifier, dans certaines circonstances, que le critère soit dichotomisé.

Résumé. Compte tenu des diverses faiblesses méthodologiques pouvant affecter une étude de validation, il faut être très prudent lors de l'appréciation du coefficient de validité obtenu auprès d'un

14. Des études solides démontrent que la relation entre les tests d'aptitudes et le rendement au travail est effectivement de nature linéaire (Coward et Sackett, 1990 ; Lubinski et Dawis, 1992). Par ailleurs, il y a d'autres postulats non moins importants à respecter, comme celui de l'homéocédasticité.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

échantillon. Il faut analyser avec soin la méthodologie de l'étude afin de détecter les principales faiblesses et ainsi tenter de savoir si le coefficient obtenu est plus faible ou plus élevé que la réalité. De façon générale, les coefficients de validité observés ont tendance à être plus faibles. D'une part, le prédicteur et le critère ne sont jamais parfaitement fidèles, ce qui entraîne une diminution de l'estimation. D'autre part, une restriction de l'échantillonnage est souvent à craindre, ce qui diminue aussi la validité observée. Même si l'erreur d'échantillonnage est toujours présente, son effet sur l'estimation de la validité est nul parce qu'elle se distribue de façon aléatoire. Quant à la contamination du critère, la direction de son effet dépend des paramètres en cause. Enfin, il est rare que le traitement statistique soit parfaitement approprié à la nature des données recueillies, ce qui entraîne souvent une sous-estimation de la validité. Bref, si l'on parvenait à corriger ces faiblesses de méthodologie, les corrélations témoigneraient de relations plus fortes qu'elles n'y paraissent.

D) Consulter les résultats des méta-analyses

De nouvelles techniques statistiques, regroupées sous le vocable de **méta-analyse**, sont maintenant utilisées pour corriger substantiellement certaines des faiblesses méthodologiques mentionnées ci-dessus. Basée sur la mise en commun de plusieurs études de validité, la méta-analyse tente d'estimer la validité réelle en compensant pour un certain nombre de faiblesses méthodologiques. Les faiblesses que la méta-analyse tentent habituellement de contrôler sont les erreurs d'échantillonnage, les erreurs dues à la restriction de l'étendue de l'échantillon et celles dues à l'infidélité du critère (Cook, 1988 ; Guion, 1991). Il faut noter que ces trois types d'erreur sont, avec l'infidélité du prédicteur, les plus importantes (Schmidt et Hunter, 1981, cité dans Lubinski et Dawis, 1992 ; Schmidt, Hunter et Pearlman, 1982). La méta-analyse est présentée plus loin à la section « Méta-analyse, généralisation de la validité et autres méthodes d'estimation ».

Quelques méta-analyses. Les résultats de plusieurs méta-analyses d'envergure ont déjà fait l'objet de publications. On en retrouve qui comparent la validité des **différents instruments** de sélection (Bobko, Roth et Potosky, 1999 ; Hunter et Hunter, 1984 ; Reilly et Chao, 1982 ; Schmitt *et al.*, 1984 ; Schmidt et Hunter, 1998), ou d'autres qui portent sur un instrument en particulier comme **l'entrevue de sélection**

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

(Huffcutt et Arthur, 1994 ; Marchese et Muchinsky, 1993 ; McDaniel *et al.*, 1994 ; Wiesner et Cronshaw, 1988 ; Wright, Lichtenfels et Pursell, 1989), le **centre d'évaluation** (Gaugler *et al.*, 1987 ; Lowenberg, Loschenkohl et Faust, 1985), les **tests de personnalité** (Barrick et Mount, 1991 ; Hogan, 1991 ; Mount et Barrick, 1995 ; Tett, Jackson et Rothstein, 1991), les **tests d'aptitudes** (Hartigan et Wigdor, 1989 ; Levine *et al.*, 1996 ; Ree et Carretta, 1998) ou **l'expérience** (Quinones, Ford et Teachout, 1995). Sans être une méta-analyse, la revue de littérature de Schippmann, Prien et Katz (1990) présente les coefficients de validité provenant de 22 études locales portant sur **l'épreuve du courrier**. Dans leur revue de littérature, Tett, Meyer et Roese (1994) résument des méta-analyses portant sur **d'autres instruments** de prédiction du rendement comme les résultats scolaires, la graphologie, les informations biographiques, etc. Des méta-analyses, présentées cette fois par type d'emploi, sont également rapportées dans le chapitre 8 de l'ouvrage de Guion (1998).

Validité concernant les tests d'aptitude mentale générale. Revenons à l'exemple des employés de production d'une usine. Ainsi, pour les tests mesurant surtout l'aptitude mentale générale ou ce que les spécialistes appellent le facteur *g*, les méta-analyses publiées par Schmidt et Hunter (1998) font état, pour l'ensemble des postes, d'une corrélation corrigée de 0,56 lorsqu'il s'agit d'un critère d'apprentissage du travail ou de formation en emploi, et de 0,51 pour un critère d'efficacité au travail. Élevées au carré, ces corrélations indiquent que l'aptitude mentale générale pourrait expliquer à elle seule chez les employés environ 31 % de l'apprentissage d'un travail et 26 % de l'efficacité de son exécution une fois en poste. Ces résultats, qui indiquent l'importance des aptitudes sur le rendement au travail, sont corroborés à la hausse par d'autres études. En effet, une méta-analyse de Levine *et al.* (1996) a donné une corrélation corrigée de 0,67 avec un critère d'apprentissage pour des emplois de métiers. À partir des données de Schmitt *et al.* (1984), couvrant un ensemble des postes variés, Ree et Carretta (1998) ont obtenu une corrélation corrigée de 0,51 pour divers critères d'efficacité une fois en poste.

Validité des tests d'aptitude mentale générale et niveau de complexité de l'emploi. Il est important de noter que les tests d'aptitude mentale générale semblent être les meilleurs prédicteurs de rendement pour toutes les tâches nécessitant une forte dose d'apprentissage, d'adaptation ou de prise de décision de la part de son

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

titulaire. De plus, indépendamment de l'apprentissage nécessité par les tâches à accomplir, ces tests demeurent parmi les meilleurs (sinon les meilleurs) prédicteurs pour presque tous les emplois, sauf les emplois très manuels ou physiques et les emplois subalternes dans la vente. Quant aux tests d'aptitudes psychomotrices, physiques et sensorielles, ils semblent plutôt convenir aux emplois manuels ou physiques qui sont moins complexes sur le plan mental. Bref, plus un emploi est exigeant du point de vue de sa complexité cognitive, plus la valeur prédictive de l'aptitude mentale générale augmente et celle des aptitudes psychomotrices diminue (Hunter, 1983, 1986 ; Lubinski et Dawis,1992).

Il ne faut donc pas perdre de vue que les corrélations mentionnées ci-dessus sont des moyennes et que la validité des tests d'aptitudes peut varier selon la complexité cognitive des emplois en cause. Ainsi, les corrélations moyennes de 0,54 (critère d'apprentissage) et de 0,45 (critère de rendement effectif) observées par Hunter et Hunter (1984) avec les tests d'aptitude mentale générale cachent des fluctuations énormes. Dans le cas d'un critère d'apprentissage, la corrélation n'est que de 0,32 pour les emplois les moins complexes, soit à peine 10,2 % d'explication, alors qu'elle grimpe à 0,76 pour les emplois les plus complexes, ce qui fait 57,8 % d'explication. Dans le cas d'un critère de rendement effectif, les corrélations varient de 0,29 à 0,61, soit de 8,4 % à 37,2 %.

Batterie générale de tests d'aptitudes (BGTA). Hartigan et Wigdor (1989) rapportent la compilation de 755 études de validité employant la *Batterie générale de tests d'aptitudes* (BGTA) et totalisant 77 141 sujets. Considérant globalement **l'aptitude mentale générale, les aptitudes perceptuo-spatiales et les aptitudes motrices**, ils obtiennent des corrélations corrigées se situant à 90 % entre 0,20 et 0,40 avec une moyenne d'environ 0,30 (corrélation non corrigée 0,22) par rapport à un critère de rendement effectif constitué de l'évaluation du superviseur. Contrairement à la démarche de Hunter et Hunter (1984), seule l'infidélité du critère a été prise en compte pour la correction des coefficients. La restriction de l'étendue de l'échantillon n'a pas été considérée, alors que l'effet de l'erreur d'échantillonnage sur l'estimation de la corrélation moyenne est négligeable, compte tenu du très grand nombre de sujets.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Validité des autres instruments de sélection. Comparativement au test d'aptitude mentale générale utilisé avec ces employés d'usine, est-ce que les autres instruments de sélection auraient pu permettre une meilleure prédiction du rendement ? Pour **l'entrevue de sélection**, les validités sont assez semblables à celles des tests d'aptitudes. Diverses méta-analyses ont été effectuées. McDaniel *et al.* (1987, cité dans Harris, 1989) obtiennent une corrélation moyenne corrigée (pour l'infidélité du critère et la restriction de l'étendue des mesures) de 0,41 (corrélation non corrigée de 0,22), alors que Wiesner et Cronshaw (1988) rapportent une corrélation semblable de 0,47 (corrélation non corrigée de 0,26). Plus récemment, Marchese et Muchinshy (1993) obtiennent, pour leur part, une corrélation corrigée de 0,38 (corrélation non corrigée de 0,27). En 1994, McDaniel et son équipe publient de nouveau et proposent une corrélation corrigée de 0,37. Enfin, Huffcutt et Arthur (1994) rapportent un coefficient corrigé de 0,37. En conclusion, selon Guion (1998), un coefficient (corrigé pour l'infidélité du critère et la restriction de l'étendue des mesures) d'environ 0,36 ou 0,37 serait actuellement une estimation raisonnable pour la validité de l'entrevue.

Cependant, les mêmes méta-analyses ont permis de constater que la validité augmente considérablement en fonction du **niveau de structure** de l'entrevue. Ainsi, si l'on ne tient compte que de l'entrevue de type très structuré, les corrélations corrigées passent respectivement à 0,45 pour McDaniel *et al.* (1987, cité dans Harris, 1989), à 0,62 pour Wiesner et Cronshaw (1988), à 0,51 pour McDaniel *et al.* (1994) et à 0,57 pour Huffcutt et Arthur (1994). Conway, Jako et Goodman (1995) estiment à environ 0,67 la limite supérieure de validité pour les entrevues très structurées et à 0,34 pour celles non structurées.

Quant au **centre d'évaluation**, qui recourt en bonne partie à des mises en situation et au jugement d'observateurs, Gaugler *et al.* (1987) ont publié une méta-analyse dans laquelle le coefficient de validité corrigé moyen est de 0,37 (corrélation non corrigée de 0,32). La validité est de 0,53 (corrélation non corrigée de 0,45) lorsque le critère est une évaluation du potentiel de gestion du candidat et de 0,36 (corrélation non corrigée de 0,31) lorsque le critère est une évaluation du rendement en emploi. Ces coefficients sont corrigés pour tenir compte de l'erreur d'échantillonnage, de la restriction de l'étendue de l'échantillon et de l'infidélité du critère. Dans une étude moins

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

récente, Schmitt *et al*. (1984) ont calculé un coefficient moyen, corrigé pour l'erreur d'échantillonnage seulement, de 0,43 lorsque le critère est une évaluation du rendement en emploi.

Voilà autant d'informations qui offrent l'occasion de mettre en perspective la validité du test d'aptitude observée chez nos 120 employés d'usine.

E) Tenir compte de l'influence d'autres facteurs sur le rendement

Une **dernière perspective**, et non des moindres, pour jauger la grandeur d'un coefficient de validité consiste à tenir compte de la multitude de variables ayant un impact sur le rendement au travail. Prédire 30 %, 20 % ou même 15 % du rendement au travail, à partir d'un seul instrument de mesure ou d'une catégorie de prédicteurs, est un résultat fort appréciable vu l'immense complexité de l'être humain et des innombrables variables qui influencent de façon dynamique son rendement dans une organisation. Lubinski et Dawis (1992) sont de cet avis et trouvent formidable, par exemple, que l'on puisse prédire plus de 25 % du rendement avec comme seuls prédicteurs des tests d'aptitudes. Comment ne pas être impressionné devant ces résultats si l'on considère que seulement sur le plan de l'individu, plusieurs facteurs sont en jeu comme ses besoins, ses valeurs, ses attitudes, ses compétences, ses aptitudes, ses traits de personnalité, etc. ? De plus, d'autres facteurs provenant de l'organisation en particulier et de l'environnement en général s'ajoutent aux facteurs individuels, et toutes ces variables interagissent sur les processus psychologiques de l'individu qui doit fournir un rendement (Pettersen et Jacob, 1992). Si ces théories sur le comportement au travail sont exactes concernant la multitude et la complexité des facteurs en cause, il n'est donc pas surprenant que la mesure d'un seul de ces facteurs ne permet que la prédiction de 15 % à 30 % du rendement au travail ; de fait, le contraire serait suspect. Tenir compte de plusieurs prédicteurs simultanément peut toutefois donner des prévisions de rendement plus appréciables.

PLUSIEURS PRÉDICTEURS ET VALIDITÉ INCRÉMENTIELLE

Aucun prédicteur ne peut à lui seul expliquer ou prédire le rendement au travail. Aussi simple que puisse être un emploi, son exécution exige un ensemble de capacités, de compétences et de dispositions

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

personnelles de toutes sortes. Pour maximiser la prédiction du rendement en emploi, il faut, comme cela se pratique dans les organisations, avoir recours à plusieurs prédicteurs pour couvrir les diverses capacités et caractéristiques des candidats. Les praticiens utilisent rarement le terme prédicteur ; ils vont plutôt employer des expressions comme profil d'exigences, critères de sélection, profil du candidat idéal ou caractéristiques recherchées. Peu importe sa désignation, chacun des prédicteurs retenus devra être aussi valide que possible par rapport au rendement dans l'emploi visé et devra ajouter quelque chose de différent par rapport aux autres prédicteurs.

Exemple. Voyons le cas de 76 employés affectés à la production dans une usine. Dans le but de faire ressortir les meilleurs tests d'aptitudes parmi plusieurs, les données suivantes ont été accumulées pendant plusieurs années en suivant un schème prédictif (voir tableau 3.2). Quatre tests psychométriques papier-crayon ont été administrés aux candidats lors de leur embauche. Le premier est l'*Examen Otis-Ottawa d'habileté mentale*, qui est un test d'aptitudes générales omnibus dont les énoncés variés relèvent des domaines verbal et éducationnel : vocabulaire, compréhension de texte, calcul, raisonnement logique, etc. Le deuxième est le *Basic Skills in Arithmetic*

Tableau 3.2
CORRÉLATIONS ENTRE QUATRE TESTS D'APTITUDES ET L'ÉVALUATION DU RENDEMENT POUR 76 EMPLOYÉS DE PRODUCTION DANS UNE USINE

	SRA	Bennett	Bêta	Rendement
Otis-Ottawa (Mentale omnibus)	0,59	0,44	0,33	0.33
SRA (Arithmétique)		0,33	0,23	0,14
Bennett (Principes mécaniques)			0,35	0,32
Bêta (Mentale non verbale)				0,31
Otis-Ottawa				0,33
Otis-Ottawa + Bêta				0,40
Otis-Ottawa + Bêta + SRA				0,40
Otis-Ottawa + Bêta + SRA + Bennett				0,40

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Test, de la firme Science Research Associates (SRA), qui mesure des habiletés de base en calcul. Le troisième est le *Test de compréhension mécanique* de Bennett portant sur la compréhension de divers principes de la physique mécanique appliqués à la vie courante. Enfin, le quatrième est le *Revised Bêta Examination* qui est un test d'aptitude générale non verbal qui n'implique aucune compréhension du langage écrit ni ne requiert de formation scolaire de base. Le critère de rendement retenu est une évaluation effectuée par le supérieur.

L'examen des corrélations par rapport à l'évaluation du rendement (dernière colonne du tableau 3.2) indique que trois des tests psychométriques ont un coefficient de validité intéressant se situant entre 0,31 et 0,33, alors que le test d'arithmétique n'affiche qu'un coefficient de 0,14 seulement. Est-ce à dire que, pour la sélection de futurs candidats, l'entreprise devrait retenir les trois tests dont la validité s'est révélée la plus élevée ? Non, du moins pas avant d'avoir vérifié jusqu'à quel point chacun des tests mesure un aspect différent et peut ainsi avoir une contribution propre à la prévision du rendement en emploi. Les corrélations entre les tests fournissent des indications à cet égard. Par exemple, la corrélation de 0,59 entre le *Otis-Ottawa* et le *SRA* indique, pour cet échantillon d'employés, que ces deux tests mesurent quelque chose en commun à 35 % (soit la corrélation de 0,59 élevé au carré), laissant une possibilité de 65 % de contribution singulière à chacun de ces deux tests. Le *Bêta* et le *SRA* semblent encore plus différents : leur corrélation de 0,23 indique seulement 5 % de redondance entre les deux (soit 0,23 élevé au carré). En revanche, la validité du *SRA* par rapport au rendement est plutôt faible.

Régression multiple. Afin de tenir compte simultanément de la validité des prédicteurs par rapport au rendement et de la redondance des prédicteurs entre eux, il faut s'en remettre à une méthode statistique appelée la **régression multiple**. Cette méthode permet de trouver la meilleure combinaison de prédicteurs pour maximiser la corrélation avec le rendement, puis de calculer cette corrélation (appelé coefficient de corrélation multiple). Sans présenter en détail cette méthode statistique de nature assez complexe pour les non-initiés, son application aux données de l'exemple a permis de trouver que la meilleure combinaison de prédicteurs dans cette situation est constituée de deux tests seulement, soit le *Otis-Ottawa* et le *Bêta*. Combinés simultanément en un score composite, les résultats à ces

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

deux tests donnent une **corrélation multiple** de 0,40 avec le rendement. Et une fois ces deux tests retenus, l'ajout du *SRA* ou du *Bennett* ne contribue pas à augmenter la prédiction du rendement : la corrélation multiple demeure 0,40. Il semble en effet que ce qui est mesuré par ces derniers n'ajoute rien à ce qui est déjà mesuré par le *Otis-Ottawa* et le *Bêta* ; il n'est donc pas utile de les retenir.

Il est important de noter que la corrélation multiple avec le rendement obtenue avec le *Otis-Ottawa* et le *Bêta* (soit 0,40) est loin de ressembler à une addition de la corrélation avec le rendement de chacun de ces tests pris individuellement (respectivement 0,33 et 0,31). Ce fait s'explique par le niveau de redondance entre ces deux tests (soit 0,33) qui diminue d'autant la contribution propre du deuxième test une fois le premier retenu. Lorsqu'il y a plusieurs prédicteurs, il faut donc procéder à la validation en considérant l'ensemble de tous les prédicteurs pris simultanément. Comme dans le présent exemple, il se peut qu'un prédicteur soit valide, mais totalement inutile parce que redondant par rapport aux prédicteurs déjà retenus.

Le coefficient de corrélation multiple peut, comme une corrélation, être élevé au carré et s'interpréter par un **pourcentage d'explication** du critère en fonction de l'ensemble des prédicteurs retenus. Par exemple, le *Otis-Ottawa* et le *Bêta* expliquent ou prédisent ensemble 16 % du rendement chez ces employés d'usine (soit 0,40 élevé au carré). Encore une fois, la prédiction peut sembler assez faible. Cependant, il ne faut pas oublier que seule l'aptitude mentale générale a été considérée, sans mesurer d'autres types de prédicteurs comme les compétences, la motivation ou la personnalité. De plus, il s'agit d'un coefficient non corrigé, qui ne compense pas pour les nombreuses limites méthodologiques présentées précédemment. Comme la corrélation, la corrélation multiple suppose une relation linéaire entre la combinaison des prédicteurs et le rendement.

Validité incrémentielle. La régression multiple est utile lorsqu'on envisage d'ajouter un prédicteur à une procédure de sélection existante ou à un ensemble de prédicteurs déjà en place. En effet, même si en soi la corrélation est très élevée entre un nouveau prédicteur et le critère de rendement (validité critériée), ce qu'il convient d'évaluer est plutôt si le prédicteur considéré ajoute quelque chose à la validité des prédicteurs déjà présents. Si le nouveau prédicteur ne

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

contribue pas à augmenter la validité de façon notable, alors sa valeur en pareille situation est discutable pour l'organisation (Guion, 1998). Cette augmentation est appelée « *incremental validity* » en anglais, expression que l'on pourrait traduire par « validité incrémentielle». Pour calculer la validité incrémentielle à l'aide de la régression multiple, il s'agit d'abord de calculer la validité totale des prédicteurs en place, à savoir la corrélation multiple entre ces prédicteurs et le critère de rendement. Ensuite, il faut introduire le nouveau prédicteur dans le calcul de la corrélation multiple. La validité incrémentielle correspond alors à l'accroissement du coefficient de corrélation multiple, si toutefois il y a accroissement.

La validité incrémentielle met en évidence la notion de complémentarité entre les divers éléments d'un ensemble. Ainsi, il ne sera pas très utile de joindre un test d'aptitude à un autre test d'aptitude, même si la validité de chacun est très élevée. Il vaut mieux chercher à compléter le profil de qualifications recherchées par des outils qui mesurent des dimensions de nature différente et qui sont complémentaires, comme des tests de personnalité ou une entrevue portant sur la motivation. Par exemple, nous savons que la validité des tests de personnalité ne semble pas très élevée comparativement aux autres outils de sélection comme les tests d'aptitudes, les examens de connaissances ou les échantillons de travail. Or, dans une étude menée auprès de vendeurs d'assurances, des mesures de personnalité se sont révélées un excellent complément à un questionnaire portant sur des données biographiques (McManus et Kelley, 1999). La validité, qui était d'environ 0,25 pour les données biographiques, est passée à plus de 0,40 avec l'ajout des mesures de personnalité. Schmidt et Hunter (1998) observent que les meilleures combinaisons d'instruments de sélection sont constituées d'une part d'un test d'aptitude mentale générale et, d'autre part, d'un échantillon de travail, d'un test d'intégrité ou d'une entrevue structurée.

Plusieurs critères. Comme pour les prédicteurs dans une régression multiple, on pourrait aussi traiter simultanément plusieurs critères de rendement au lieu d'un seul à la fois, le rendement au travail étant constitué de plusieurs dimensions différentes. De plus, comme pour les prédicteurs encore, les différentes dimensions du rendement peuvent être corrélées entre elles et comporter une certaine redondance. Traditionnellement, cette information n'était pas prise en compte dans les études de validité qui ne retenaient qu'un seul critère

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

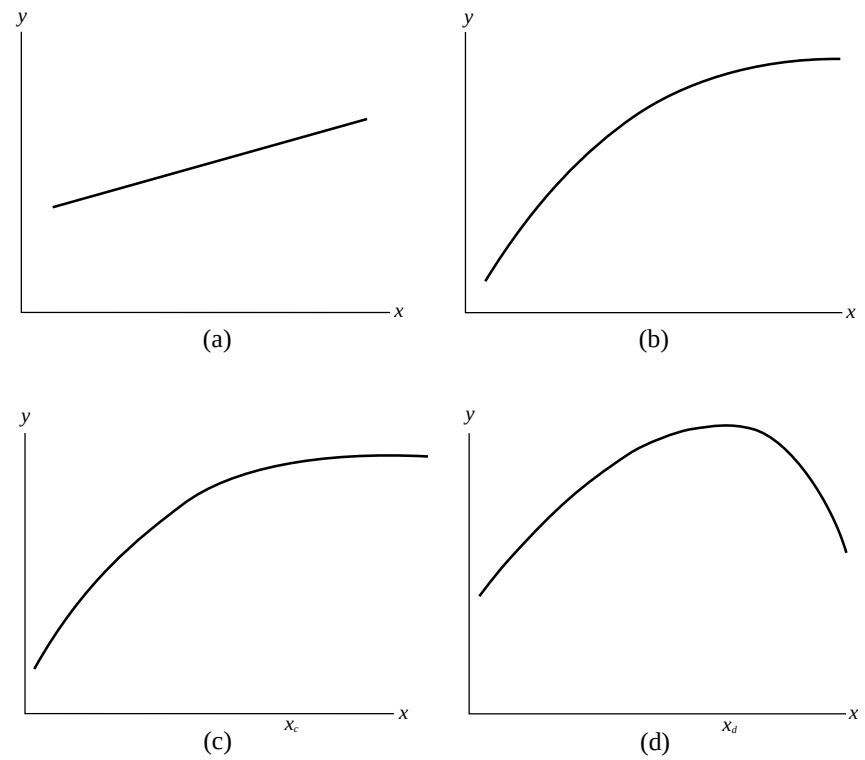
Tous droits de reproduction, de traduction ou d'adaptation réservés

de rendement ou, s'il y en avait plusieurs, les traitaient séparément. Afin de tenir compte des multiples dimensions du critère et ainsi mieux évaluer la validité des prédicteurs, il faudrait mettre en corrélation simultanément les prédicteurs retenus et plusieurs dimensions du rendement (McHenry *et al.*, 1990 ; Murphy et Shiarella, 1997).

RELATION NON LINÉAIRE

Le coefficient de corrélation n'est pas approprié pour rendre compte de la relation entre le prédicteur et le critère lorsque cette relation n'est pas de type linéaire. En sélection, on peut retrouver différentes formes de relation entre le prédicteur (axe des *x*) et le critère de rendement (axe des *y*). La figure 3.8, tirée de Guion (1991), illustre quatre formes caractéristiques de relation. La **relation (*a*)** est une

Figure 3.8
QUATRE FORMES CARACTÉRISTIQUES DE RELATION PRÉDICTEUR-CRITÈRE



© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

relation linéaire. Dans ce cas, toute variation au prédicteur, peu importe le niveau sur l'axe des *x*, est accompagnée par une variation constante du critère. Par exemple, une variation donnée du prédicteur entraîne la même variation au critère de rendement, que la différence au prédicteur soit observée chez les candidats faibles, moyens ou forts. La **relation (*b*)** est de forme non linéaire monotonique. Une différence au prédicteur pour des candidats faibles est accompagnée d'une différence plus forte au critère que cette même différence au prédicteur pour des candidats forts. La **forme (*c*)** est semblable à la forme (*b*) jusqu'au point X_c, après quoi il n'y a plus de relation entre le prédicteur et le critère. Quant à la **forme (*d*)**, elle indique qu'une augmentation du prédicteur s'accompagne d'une augmentation du critère de rendement jusqu'à un certain point X_d du prédicteur. Passé ce point, la relation s'inverse de sorte qu'une augmentation du prédicteur est associée à une diminution du critère de rendement.

Lorsque la relation avec le critère n'est pas linéaire, qu'advient-il de la validité d'un instrument de mesure ? D'abord, la validité ne dépend pas de la forme de la relation, mais de l'existence de cette relation. S'il y a une relation, linéaire ou non, il y a validité et l'instrument de mesure peut être utilisé. Dans le cas des formes (*a*) et (*b*) par exemple, la sélection des candidats pourrait se faire en retenant les candidats les plus forts, jusqu'à ce que le nombre de personnes requises soit obtenu. Dans le cas d'une relation comme celle qui existe dans la situation (*c*), la sélection des candidats pourrait se faire à l'aide d'une note de passage minimale au prédicteur, alors que dans la situation (*d*), on pourrait avoir deux notes de passage, une minimale et une maximale.

Relation entre les aptitudes et le rendement. La nature de la relation qui existe entre le rendement et les divers types de prédicteurs dépend de plusieurs facteurs, ce qui fait que l'on ignore si elle est de nature linéaire ou non. Cependant, une étude récente et d'envergure étaie solidement une **tendance générale vers une relation linéaire** entre les tests d'aptitudes et le rendement au travail, et cela, à tous les niveaux d'aptitude (Coward et Sackett, 1990). Citant les travaux de Dawes (Dawes, 1979 ; Dawes et Corrigan, 1974 ; Dawes, Faust et Meehl, 1989), Lubinski et Dawis (1992) sont aussi de cet avis.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

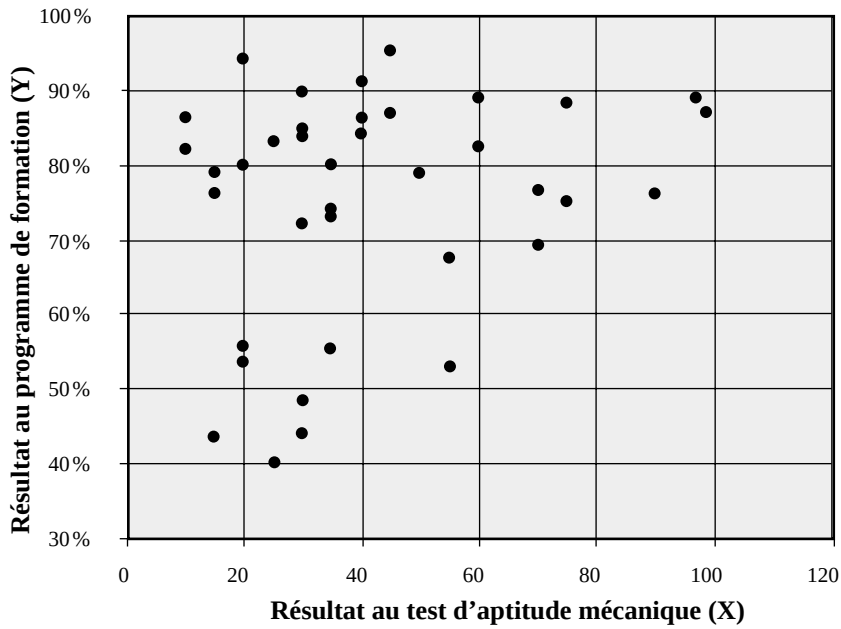
Tous droits de reproduction, de traduction ou d'adaptation réservés

Occasionnellement, il peut arriver qu'une étude locale fasse **exception** à cette tendance et mette en évidence une relation non linéaire. Lors d'un cours sur l'utilisation des tests psychométriques, un participant se plaignait d'éprouver des difficultés avec ses employés depuis que le service du personnel utilisait un test pour en faire la sélection ; il se plaignait, entre autres, de difficultés à faire respecter l'autorité et d'un niveau très élevé de plaintes et de griefs de toutes sortes. Après discussion, nous avons appris que le test en question était un test d'aptitude mentale générale et que seuls les candidats ayant obtenu les scores les plus élevés étaient sélectionnés alors que les tâches étaient extrêmement simples et ennuyeuses. Bref, nous étions vraisemblablement en présence d'une situation de surqualification ; les employés choisis ne pouvaient pas exploiter pleinement leur potentiel. Il y avait alors risque de démotivation et de frustration, risque attesté par les symptômes manifestés par les employés. La relation entre le test d'aptitude et le rendement dans cette situation a tendance à prendre la forme (*d*) (figure 3.8).

Voyons un autre exemple d'étude locale où, cette fois, la nature de la relation ne semble pas constante pour les divers niveaux de résultats au prédicteur. Lors d'un recrutement à l'interne, on a fait passer un test d'aptitude mécanique à 38 employés retenus pour suivre un programme de formation technique. Au cours de cette période de formation d'une durée d'une semaine, les candidats ont été évalués à plusieurs reprises par des examens. La figure 3.9 présente le diagramme de dispersion pour ces 38 employés entre leur score au test d'aptitude mécanique et la moyenne globale obtenue aux examens. On voit immédiatement que la répartition du nuage de points n'est ni linéaire ni aléatoire : il n'y a pratiquement pas d'éléments dans la partie inférieure droite. Les candidats dont les scores au test sont les plus élevés ne semblent pas avoir éprouvé de difficultés aux examens techniques administrés en cours de formation ; la majorité ont des résultats assez élevés, supérieurs à 70 %. En revanche, la situation est très différente pour les candidats dont les résultats au test d'aptitude mécanique sont plus faibles ; certains ont bien réussi les examens alors que d'autres ont éprouvé des difficultés importantes. Pour ces candidats plus faibles au test, il ne semble plus y avoir de relation entre le test et la réussite au programme de formation.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 3.9
DIAGRAMME DE DISPERSION ENTRE UN TEST D'APTITUDE MÉCANIQUE ET LE RÉSULTAT À UN PROGRAMME DE FORMATION POUR 39 EMPLOYÉS D'UNE USINE



Abstraction faite de la très petite taille de l'échantillon, on pourrait affirmer qu'il existe une certaine validité pour ce test et qu'il pourrait servir à identifier les candidats ayant le plus de chances de réussir leur formation technique. Cependant, cette relation n'est pas uniforme pour les divers niveaux de résultats au prédicteur. Par conséquent, la corrélation de 0,24, compilée à partir de ces données, sous-estime l'ampleur réelle d'une telle relation. Il faut se référer à des ouvrages de statistiques pour le traitement de ce type de relations non linéaires (voir chapitre 7, section « Coefficients de corrélation et de détermination »).

UNE VARIABLE NOMINALE ET UNE D'INTERVALLE (COMBINAISONS 3 ET 7)

Contrairement à la combinaison précédente où le prédicteur et le critère sont de niveau d'intervalle (combinaison 9), il arrive que l'une ou l'autre des variables soit de niveau nominal (voir figure 3.4).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Ainsi le prédicteur peut être nominal (combinaison 7) : par exemple, avoir un permis de conduire ou non, détenir au moins trois ans d'expérience en comptabilité, etc. Dans l'autre cas, c'est le critère qui est nominal (combinaison 3) : par exemple, réussir le programme d'entraînement à la tâche, avoir reçu une promotion, etc.

Comparaison des moyennes. Dans ces situations, on peut voir s'il y a une relation entre le prédicteur et le critère en comparant les moyennes des groupes formés par les différentes valeurs de la variable nominale. Si c'est le prédicteur qui est nominal, on compile les moyennes au critère pour chacun des groupes distincts au prédicteur. Réciproquement, si c'est le critère qui est de niveau nominal, on compile les moyennes au prédicteur.

Prenons l'exemple des étudiants en sciences comptables qui se sont présentés aux examens de l'Ordre des comptables agréés (voir chapitre 7). Considérons les résultats aux examens de l'Ordre non pas en fonction des scores bruts de 150 à 340 (échelle de niveau d'intervalle), mais en fonction du succès ou de l'échec (échelle pouvant être considérée comme de niveau nominal). À partir des données recueillies au tableau 7.7, comment savoir s'il y a une relation entre la moyenne cumulative (prédicteur) et le succès aux examens de l'Ordre (critère) ? Pour répondre à cette question, il suffit de compiler la moyenne cumulative universitaire moyenne pour les finissants qui ont réussi les examens de l'Ordre, puis de la comparer à la moyenne cumulative moyenne de ceux qui ont échoué. Cette compilation, présentée au tableau 3.3, révèle que les étudiants ayant réussi leur examen ont globalement une moyenne cumulative de 3,054, comparativement à une moyenne de 2,432 pour ceux qui ont échoué, soit une différence de 0,622. Les étudiants qui réussissent leur examen de l'Ordre des comptables agréés réussissent également mieux leurs études universitaires, ce qui prouve l'existence d'une relation entre la moyenne cumulative et le succès aux examens des comptables agréés[15].

15. Rappelons que, lorsqu'une variable est continue, il n'est pas recommandé d'en réduire la distribution à quelques points (p. ex., « succès-échec » ou « fort-moyen-faible »), comme cela se fait actuellement avec les résultats aux examens des comptables agréés, car cette pratique diminue la puissance des tests statistiques (voir section « Considérer les faiblesses méthodologiques »).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 3.3
COMPARAISON DES MOYENNES CUMULATIVES DE 46 FINISSANTS AU BACCALAURÉAT EN SCIENCES COMPTABLES (1980)

	Nombre	Moyenne	Écart type
Ont réussi	24 (52 %)	3,054	0,394
Ont échoué	22 (48 %)	2,432	0,314
Total	46 (100 %)	2,750	0,474

Tests statistiques. Est-ce que cette différence de 0,622 est d'une ampleur appréciable ? Peut-elle être simplement le fruit du hasard ? Pour répondre à ces nouvelles questions, il faut de nouveau faire appel aux méthodes statistiques et à ses différents tests de signification. Dans l'exemple qui nous occupe, il serait approprié d'effectuer un test *t* de Student, car la variable nominale (résultats aux examens de l'Ordre) ne prend que deux valeurs (succès ou échec). Un tel test donnerait une valeur *t* égale à 5,94, qu'il faut comparer aux valeurs *t* d'une table de signification (voir tableau A.3, appendice A). Dans cet exemple, le degré de liberté est égal à 44 (soit le nombre de sujets moins deux). La table ne comportant pas tous les degrés de liberté, il faut s'en remettre au cas le plus proche qui est 40 de degré de liberté. Les valeurs *t* correspondantes (pour un test unilatéral) sont alors de 1,684 pour une probabilité de 0,05 et de 2,423 pour une probabilité de 0,01. La valeur *t* compilée de 6,43 étant nettement au-dessus des valeurs de la table, il est donc permis d'affirmer, avec un risque d'erreur inférieur à 1 %, que la différence observée dans les moyennes universitaires entre ceux qui réussissent et ceux qui échouent leurs examens des comptables agréés est réelle et n'est pas le fruit du hasard.

Si la variable nominale avait pris plus de deux valeurs (soit forts, moyens et faibles), ce qui impliquerait de compiler plus de deux moyennes, il aurait fallu s'en remettre à un test F d'analyse de la variance. Il existe d'autres méthodes et tests statistiques qui s'appliquent à la présente combinaison d'une variable nominale et d'une variable d'intervalle. Pour en savoir plus, il faut consulter un ouvrage en statistique.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

PRÉDICTEUR ET CRITÈRE DE NIVEAU NOMINAL (COMBINAISON 1)

Table de contingences. Il arrive que le prédicteur et le critère soient tous les deux de niveau nominal (combinaison 1). La méthode la plus utilisée pour analyser la relation est alors de commencer par la compilation d'une table de contingences. L'exemple fictif suivant permettra d'illustrer cette méthode. Une entreprise constate qu'elle a un fort taux de roulement parmi ses nouveaux vendeurs. Environ 60 % des personnes embauchées comme vendeurs quittent l'entreprise avant d'avoir terminé leur première année. La directrice des ressources humaines croit que les longues heures de travail, réparties selon un horaire très irrégulier, entrent peut-être en conflit avec les exigences d'une vie familiale chargée, comme c'est le cas pour les parents vivant seuls et ayant la charge d'enfants. Elle veut donc vérifier si, effectivement, il y a une relation entre le taux de roulement et le niveau d'exigences familiales. Ainsi, la directrice des ressources humaines prend note, pour chacune des personnes engagées comme vendeur au cours des 10 dernières années, de la durée de son emploi avec l'entreprise, de son statut familial et du nombre d'enfants à charge. Avec ces données, elle compile une table de contingences (voir tableau 3.4).

Tableau 3.4
TAUX DE ROULEMENT DES NOUVEAUX VENDEURS EN FONCTION DE LEUR STATUT FAMILIAL

	Statut familial			
Durée à l'emploi	**Sans enfant, avec ou sans conjoint**	**Avec enfant(s), avec conjoint**	**Avec enfant(s) sans conjoint**	**Total**
A quitté avant un an	7 (27 %)*	26 (50 %)	72 (74 %)	105 (60 %)
Est demeuré plus d'un an	19 (73 %)	26 (50 %)	25 (26 %)	70 (40 %)
Total	26 (15 %)	52 (30 %)	97 (55 %)	175 (100 %)

Note : Les pourcentages des cellules sont calculés par rapport au nombre total de personnes par colonne. Au bout des lignes et au pied des colonnes, les pourcentages sont calculés par rapport au nombre total.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Il semble bien y avoir une relation entre le fait d'abandonner son emploi avant un an et les exigences de la vie familiale. En effet, 74 % des personnes sans conjoint mais ayant des enfants à leur charge, vraisemblablement les personnes ayant les plus lourdes tâches familiales, ont quitté leur emploi avant un an. Chez les personnes avec enfants mais aussi avec conjoint, le taux de roulement baisse à 50 %. Finalement, seulement 27 % des personnes n'ayant pas d'enfant ont quitté.

Tests statistiques. Comme pour les autres combinaisons, il existe différents tests statistiques pour quantifier la relation prédicteur-critère observée dans une table de contingence ; le plus utilisé est le khi carré (χ^2).

RÉSUMÉ

L'étude de la relation entre le prédicteur et le critère se fait en deux étapes. D'abord, les données sont présentées de manière à pouvoir visualiser la relation. Ensuite, divers indices statistiques sont compilés afin de quantifier la grandeur de cette relation ; ces indices peuvent être comparés à des valeurs seuils pour vérifier si la grandeur de la relation observée est réelle ou simplement due au hasard. Le tableau 3.5 résume ces deux étapes pour chacune des combinaisons étudiées. Il convient de rappeler que les autres combinaisons met-

Tableau 3.5
RÉSUMÉ DE L'ANALYSE EN DEUX ÉTAPES DE LA RELATION PRÉDICTEUR-CRITÈRE POUR CHACUNE DES COMBINAISONS ÉTUDIÉES

Combinaison	Première étape : visualiser la relation	Deuxième étape : quantifier la relation et test de signification
Combinaison 9 : prédicteur et critère de niveau d'intervalle	Diagramme de dispersion	Coefficient de corrélation
Combinaisons 3 et 7 : une variable nominale et une variable d'intervalle	Comparaisons des moyennes	Test *t* de Student, analyse de la variance, etc.
Combinaison 1 : prédicteur et critère de niveau nominal	Table de contingences	Khi carré, etc.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

tant en cause une variable de **niveau ordinal** sont souvent traitées comme si ces variables étaient de niveau d'intervalle. Même si c'est habituellement le cas, nous ne pouvons trop insister sur les limites que peut avoir un tel traitement statistique et, conséquemment, nous invitons le lecteur à se documenter davantage.

DOUBLE VALIDATION

Chaque évaluation, chaque mesure, chaque statistique contient une portion d'erreur ; une partie de cette erreur est aléatoire, l'autre systématique (voir chapitre 4). Par exemple, pour la portion aléatoire, on a vu plus tôt dans ce chapitre que **l'erreur d'échantillonnage** est inévitable lorsqu'une étude porte sur un échantillon plutôt que sur la population entière ; et plus l'échantillon est petit, plus ce risque d'erreur, qui est le simple fait du hasard, augmente. On sait également que les **mesures du prédicteur et du critère**, dont la fidélité n'est jamais parfaite, sont aussi une grande source d'erreur aléatoire. Quant à la portion d'erreur systématique, elle provient de divers **biais** comme le choix d'un échantillon de candidats à partir d'un seul sous-groupe de la population cible, d'une mesure du critère qui avantagerait indûment une partie de l'échantillon, etc.

Étant donné l'importance éventuelle de ces erreurs, il est déconseillé de se fier à une seule étude pour estimer la validité. Pour s'assurer de la stabilité de la relation entre le prédicteur et le critère, la sagesse exige d'effectuer une **double validation**, c'est-à-dire contrôler les résultats auprès d'un nouvel échantillon de candidats (Dunnette, 1966). La nécessité de cette vérification est d'autant plus grande lorsque l'étude de validation repose sur un petit échantillon ou sur une approche statistique comme la régression multiple. La double validation importe également en raison des changements réels qui peuvent survenir dans la relation prédicteur et critère. Par exemple, des changements dans les tâches ou les conditions de travail, comme lors de l'implantation d'une nouvelle technologie, peuvent modifier les exigences à l'égard des employés. La composition de la main-d'œuvre peut évoluer et faire en sorte que se modifie la pertinence d'une exigence (p. ex., une main-d'œuvre moins expérimentée peut accroître l'importance de la capacité d'apprentissage). En résumé, la double validation est un moyen de minimiser les erreurs et de

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

suivre l'évolution réelle de la relation entre le prédicteur et le critère. La double validation, qui n'est en fait qu'un recommencement du processus de validation, devrait logiquement s'effectuer de manière cyclique (voir figure 3.1).

Répétition et contre-validation. Nous allons maintenant aborder un aspect technique dont la compréhension peut exiger des connaissances un peu plus approfondies en statistique. Traditionnellement, les auteurs ont eu tendance à utiliser la même expression, soit « contre-validation » (*cross validation*), pour désigner en fait deux procédures distinctes de double validation. La première, que Guion (1998) propose de nommer simplement « **répétition** » (*replication*), consiste effectivement à « […] répéter l'étude originale, avec ou sans modifications systématiques aux instruments de mesures ou aux procédures, afin de voir si une étude indépendante produit des résultats similaires » (p. 353).

La deuxième, soit la vraie procédure de « **contre-validation** », ne devrait s'appliquer que lorsque les coefficients sont obtenus par régression multiple auprès d'un échantillon, puis comparés à ceux d'un autre échantillon. Les résultats étant spécifiques à l'échantillon, la régression multiple a tendance à surestimer la corrélation multiple entre les prédicteurs et le critère (*statistical overfitting*). Plus les prédicteurs sont nombreux et plus l'échantillon est petit, plus le risque de surestimation augmente (Schmitt et Chan, 1998).

Comme pour la répétition, la contre-validation devrait idéalement être effectuée à l'aide d'un nouvel échantillon, indépendant du premier. En effet, il existe une approche qui consiste à ne recourir qu'à un seul échantillon que l'on scinde en deux, puis à compiler les coefficients de validité séparément pour chaque moitié d'échantillon. Cette approche vaut mieux que de ne rien faire du tout de l'avis de Guion (1998), mais il est clair qu'il ne s'agit pas vraiment de deux échantillons indépendants. S'il existait un biais lié à l'échantillon original, il va se retrouver vraisemblablement dans chacun des deux sous-échantillons. Lorsqu'il est impossible de recourir à un deuxième échantillon indépendant, il vaut peut-être mieux compenser la surestimation des coefficients obtenus par régression multiple en appliquant une formule appropriée (voir Cattin, 1980).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Efficacité douteuse de la double validation. La double validation, que ce soit une répétition ou une contre-validation, est une technique simple, mais son efficacité est plus apparente que réelle. Voyons un exemple basé sur des données réelles. L'étude de validation portant sur la moyenne universitaire cumulative comme prédicteur des résultats aux examens de l'Ordre des comptables agréés a donné une corrélation de 0,65 pour un groupe de finissants de 1979. Répétée rigoureusement en 1980, la seconde étude a permis d'obtenir une corrélation de 0,77. Au regard du pourcentage de prédiction, c'est une augmentation de 17 points ($0{,}65^2 = 42{,}3\,\%$, $0{,}77^2 = 59{,}3\,\%$) entre les deux années ! Est-ce que cette augmentation substantielle est simplement l'effet d'erreurs provenant de l'échantillonnage, de l'infidélité des mesures ou d'autres faiblesses méthodologiques ? Si tel est le cas, on n'a qu'à combiner les deux corrélations en une moyenne et estimer que la validité se situe autour de 0,71. Mais que se passe-t-il si cette augmentation témoigne plutôt d'une différence réelle entre les deux situations de prédiction, d'un changement qui serait survenu entre les deux études, tel un changement dans les conditions de réussir aux examens de l'Ordre ou dans le contenu du programme universitaire ? Il faudrait alors se fier au résultat de l'étude la plus récente. Or, auprès des finissants de 1985, une autre étude a donné une corrélation de 0,68, soit 46,2 % de prédiction. Est-ce l'indication d'un nouveau changement réel ou plutôt la preuve qu'il ne s'agit en fait que d'erreurs aléatoires ? Quelle est la meilleure estimation de la validité de la moyenne universitaire cumulative comme prédicteur des résultats aux examens de l'Ordre des comptables agréés ? Malgré trois études rigoureuses, on doit admettre en toute humilité qu'on ne le sait pas.

Une double validation vaut mieux qu'une seule étude ; mais de là à être certain du coefficient de validité obtenu, il y a un grand pas. Même en répétant plusieurs fois l'étude de validation, avec ou sans modifications à la procédure ou aux instruments de mesure, il est difficile de savoir si la différence entre les coefficients de validité ainsi obtenus est le résultat d'erreurs ou traduit une différence réelle entre les situations de chaque étude. Il faut se rendre à l'évidence qu'une étude locale de validation, même accompagnée d'une double validation, comporte des limites importantes lorsqu'il s'agit d'estimer la

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

validité dans une situation de prédiction[16]. Une solution plus satisfaisante à ce problème est fournie par la méta-analyse et la généralisation de la validité.

MÉTA-ANALYSE, GÉNÉRALISATION DE LA VALIDITÉ ET AUTRES MÉTHODES DE VALIDATION

Méta-analyse. Au cours des années 1970, plusieurs auteurs se sont attaqués sérieusement au problème d'estimation de la validité en concevant une nouvelle approche, la méta-analyse. Cette approche consiste à combiner quantitativement les résultats de plusieurs études locales afin de déterminer notamment 1) la relation réelle entre deux variables et 2) la part de variation observée entre les études qui est imputable au hasard et autres faiblesses méthodologiques (appelés alors artefacts statistiques) et celle imputable aux différences réelles existant entre les études (présence de variables modératrices) [Fried et Ager, 1998 ; Tett, Meyer et Roese, 1994]. Il peut être intéressant de savoir que l'origine de la méta-analyse est attribuée aux travaux de Glass et Smith en psychologie clinique (1976, cité dans Hunt, 1997) et que la méthode s'est très rapidement répandue à d'autres secteurs de recherche, notamment à la médecine, à la physique et à l'éducation (Hartigan et Wigdor, 1989 ; Hunt, 1997). Son approche quantitative est probablement l'un des aspects qui ont contribué à son succès presque instantané, même si Guion (1998) nous rappelle que la méta-analyse recèle une part de subjectivité dans le choix et la codification des études.

Généralisation de la validité. Des différentes approches qui existent en méta-analyse, celle connue sous le nom de « généralisation de la validité » est la plus utilisée en sélection du personnel ; elle a été introduite par les travaux de Schmidt et Hunter (1977, cité dans Guion, 1998). La logique à la base de la généralisation de la validité est très simple. D'abord, il s'agit de rassembler plusieurs coefficients de validité obtenus dans des études indépendantes, mais portant sur la même problématique (p. ex., la validité des tests d'aptitude mentale

16. La situation de prédiction est définie par l'objectif du programme de validation. Dans cet exemple, la situation de prédiction est le succès aux examens de l'Ordre pour les finissants en comptabilité, anciens et futurs, de cette même université.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

générale à prédire le rendement en emploi). Les coefficients ainsi obtenus ne sont pas tous identiques et varient d'une étude à l'autre. D'un point de vue logique, cette différence entre les coefficients de validité ne peut avoir que deux significations. Ou bien ces différences sont artificielles et ne font que signaler la présence d'erreurs causées par des faiblesses méthodologiques (p. ex., erreurs d'échantillonnage) ; on parle alors d'**artefacts** statistiques. Ou bien ces différences sont réelles et traduisent la spécificité de certaines situations ou de certaines études (p. ex., les exigences de réussite ont changé au cours des années ou encore chaque emploi a des exigences différentes pour ce qui est des aptitudes mentales) ; il y a alors présence de **variables modératrices** qui influent sur les coefficients de validité.

Par des calculs appropriés, il est possible d'évaluer la portion de cette variation des coefficients qui dépend des artefacts d'origine méthodologique, puis d'inférer par simple soustraction la portion qui dépend de la spécificité des études. Si une faible portion de variation dépend des artefacts, on conclut à la présence de variables modératrices et les coefficients de validité sont considérés spécifiques à chaque situation. En revanche, si la grande majorité de la variation (disons 75 % et plus) est attribuable aux artefacts, on en conclut que les différences observées sont artificielles et que la validité est généralisable à l'ensemble des situations étudiées. On peut alors calculer le coefficient de validité général, ou moyen, pour ces situations, tout en éliminant l'effet néfaste des artefacts considérés. Ce coefficient, appelé **coefficient corrigé de validité**, est une estimation de la validité réelle, c'est-à-dire la validité qui serait obtenue si les erreurs méthodologiques considérées étaient absentes.

Il va sans dire que la qualité de l'estimation dépendra du nombre d'artefacts qu'il aura été possible de corriger. Les artefacts pris en considération et qui sont les plus importants comprennent habituellement l'erreur l'échantillonnage, l'infidélité de la mesure du critère et la restriction de l'étendue des mesures (Cook, 1988 ; Guion, 1991 ; Schmitt et Chan, 1998). Excepté avec des échantillons très grands, l'erreur d'échantillonnage serait la plus importante dans la plupart des cas (Schmidt, Hunter et Pearlman, 1982) et pourrait être responsable de la moitié de la variation dans les coefficients de validité entre les études (Lubinski et Dawis, 1992). Or, la taille typique des études de validation est relativement petite (Catano *et al.*, 1997), autour de 68 sujets selon le recensement de Lent *et al.* (1971), et on sait que

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

plus l'échantillon est petit, plus l'erreur d'échantillonnage peut être grande. Bref, en plus de reposer sur plusieurs, sinon sur presque toutes les études de validation recensées pour une situation donnée, le coefficient corrigé de validité permet de contrôler les principaux artefacts méthodologiques.

IMPACTS DES MÉTA-ANALYSES SUR LA GESTION DES RESSOURCES HUMAINES

Les résultats obtenus par cette approche, regroupés indistinctement dans la documentation sous les vocables de méta-analyse ou de généralisation de la validité, ont littéralement révolutionné les pratiques de validation (Guion, 1991 ; Hoffman et McPhail, 1998) ; certains auteurs n'hésitent pas à parler d'une ère nouvelle, celle de la gestion scientifique des ressources humaines (Smith, Gregg et Andrews, 1989). Tous les instruments de sélection ont fait l'objet à ce jour de méta-analyses, particulièrement les tests psychométriques d'aptitudes et de personnalité ainsi que l'entrevue. Même si les chercheurs sont encore à l'œuvre, les retombées de ces études sont considérables pour la pratique (Cook, 1988 ; Guion, 1991 ; Lubinski et Dawis, 1992).

Validité réelle plus élevée. Premièrement, grâce au calcul du coefficient corrigé de validité, on a pu réaliser que la validité réelle des différents outils de sélection est plus élevée que l'on croyait (voir section « Consulter les résultats des méta-analyses »). Par exemple, la validité des tests d'aptitudes est probablement autour de 0,50, ce qui est bien supérieur aux sempiternelles corrélations observées de 0,20 à 0,30 dans les diverses études locales de validité. Cela représente une véritable renaissance des tests psychométriques, passablement délaissés au cours des années 1960 et 1970. Et que dire de l'entrevue dont la validité s'est vue réhabiliter avec des coefficients autour de 0,37 et pouvant même dépasser 0,50 dans le cas d'entrevue très structurée. La validité élevée du centre d'évaluation, des tests situationnels, des examens de connaissances et autres échantillons de travail a été confirmée avec des coefficients souvent supérieurs à 0,40. Pas de chance pour la graphologie : sa validité demeure quasiment nulle.

Affaiblissement de l'hypothèse de la spécificité. Deuxièmement, l'hypothèse de la spécificité, surtout pour les tests d'aptitudes, est fortement affaiblie. En effet, selon la croyance traditionnelle en

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

psychométrie, un test devait être validé dans sa situation spécifique d'utilisation en raison des exigences de l'emploi présumées uniques pour chacune des situations. Autrement dit, les coefficients de validité obtenus dans une situation n'étaient pas exportables dans une autre situation. Par exemple, un test d'aptitudes verbales ayant fait ses preuves pour des vendeurs d'assurances dans la région de Québec ne pourrait pas être utilisé pour les mêmes fins dans la région de Montréal, sans reprendre de nouvelles études de validité. Haccoun (voir Tziner, Jeanrie et Cusson, 1993) rappelle à juste titre que l'expérience soutenait cette pratique. À plusieurs reprises, on avait observé que la validité d'un instrument pouvait fluctuer de façon importante d'une situation à l'autre. Il s'agit de se rappeler l'exemple des examens à l'Ordre des comptables agréés ou de consulter les impressionnantes compilations effectuées par Ghiselli (1966, 1973) pour s'en convaincre.

Or, il semble de plus en plus établi que ces variations de coefficients de validité observées d'une étude à l'autre étaient moins souvent le reflet de changements réels dans la situation que le fruit d'erreurs, d'artefacts méthodologiques, dont les principaux étaient l'erreur d'échantillonnage, l'infidélité de la mesure du critère et la restriction de l'étendue des données. La validité des divers outils de sélection ne serait donc pas aussi spécifique que les résultats le laissaient croire. Certains auteurs estiment cependant que les techniques actuelles de méta-analyse exagèrent la généralisation de la validité au détriment de l'hypothèse de la spécificité (Erez, Bloom et Wells, 1996). C'est un débat à suivre.

La validité peut être généralisée. Ainsi, les études locales de validité nous avaient partiellement induits en erreur : la validité des instruments de sélection est moins spécifique et, en moyenne, plus élevée que ce que l'on croyait, ce qui entraîne une troisième retombée. En effet, on peut dorénavant envisager de généraliser la validité d'un instrument de sélection à toute une classe de situations hautement similaires. L'étude locale de validation n'est plus jugée aussi essentielle qu'avant (Lubinski et Dawis, 1992) ; conséquemment, un employeur peut utiliser directement un test, sans étude de validation, s'il est en mesure de démontrer que ce test a donné de bonnes prédictions dans des situations similaires. Par exemple, si les résultats d'une méta-analyse indiquent que les tests d'aptitude mentale générale ont une bonne validité prédictive, disons 0,40, pour l'ensemble

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

des emplois de secrétariat, alors on peut généraliser qu'un test de cette catégorie aura une validité sensiblement comparable pour un poste particulier appartenant à cette classe.

Fondamentalement, cette généralisation à partir des résultats de méta-analyses n'est justifiable que si la situation locale de sélection est **similaire** aux situations étudiées par la méta-analyse. En termes plus techniques, Guion (1998) précise que les variables et les caractéristiques des études rassemblées dans la méta-analyse doivent correspondre, sinon les inclure, à celles de la situation spécifique envisagée. Il soumet ensuite quelques aspects à considérer pour évaluer jusqu'à quel point la généralisation est adéquate.

En 1985, l'American Educational Research Association, l'American Psychological Association et le National Council on Measurement in Education recommandaient de recourir à la généralisation de la validité seulement lorsqu'on ne disposait pas d'une étude de validation locale adéquate (article 1.16). En 1999, ces mêmes associations ont changé leur position : elles reconnaissent qu'il n'est plus nécessaire de mener une étude locale si la littérature fournit déjà des résultats d'études suffisamment abondants et consistants. Lorsque ces résultats sont particulièrement probants, ils recommandent même de considérer avec prudence des résultats contradictoires qui seraient obtenus par une étude locale. Guion (1998) va plus loin : il conseille de faire une étude locale seulement lorsqu'il n'y a pas de méta-analyse appropriée à la situation en cause. Comme d'autres (Schmidt et Hunter, 1980 ; Schmidt *et al.*, 1985), Guion considère que l'étude locale de validation n'est pas un moyen très puissant pour estimer la validité et que les méta-analyses sont plus fiables.

Alors, à quoi bon les études locales de validation ? Si la méta-analyse est un moyen plus fiable pour estimer la validité d'un instrument de mesure, à quoi servent alors les études locales de validation ? Premièrement, il y a des situations pour lesquelles il **n'existe pas** de méta-analyse pertinente ; il faut donc effectuer sa propre étude locale. Deuxièmement, s'il n'y a plus d'études locales, la méta-analyse ne pourra **plus être appliquée**. La méta-analyse est une technique qui s'appuie justement sur la compilation de plusieurs études locales réalisées au fil des ans ; c'est en quelque sorte une méthode pour agréger les études spécifiques et en faire le bilan. Par exemple, de nouvelles études locales seraient actuellement nécessaires pour faire

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

progresser les connaissances sur les facteurs (appelés variables modératrices) qui affectent la relation prédicteur-critère (Schuler et Guldin, 1991). Troisièmement, il ne faut pas tomber dans l'**excès contraire**. Même si la validité d'un instrument de mesure peut se généraliser à l'ensemble d'une situation, une partie de cette validité peut demeurer spécifique à un contexte, à un groupe ou à une tâche en particulier. Selon Guion (1991, 1998), il est antiscientifique de généraliser sans poursuivre la recherche des exceptions, ajoutant du même souffle que la méta-analyse a aussi ses faiblesses. Quatrièmement, une étude locale fournit des données nécessaires à l'établissement d'une **note de passage** et à sa justification. Cinquièmement, même si les **tribunaux** acceptent la généralisation de la validité, ils continuent d'accorder une grande importance aux études locales (Schmidt *et al.*, 1985, cité dans Cook, 1988)[17].

Deux démarches de validation complémentaires. Jusqu'ici, deux démarches sont disponibles pour estimer la validité critériée d'un instrument de mesure : l'étude locale de validation, incluant si possible une étape de double validation, et la consultation des méta-analyses pertinentes. Idéalement, les deux démarches devraient être entreprises parce qu'elles se complètent mutuellement. Le processus de validation comporte ses avantages propres, dont certains sont particulièrement importants en pratique (justifier la note de passage, se défendre devant les tribunaux, tenir compte d'un changement réel dans la situation, etc.). Cependant, comme il y a toujours possibilité d'erreurs méthodologiques lors d'un tel processus (restriction de l'étendue de l'échantillon, fidélité de la mesure du critère, fidélité de la mesure du prédicteur, erreur d'échantillonnage, etc.), le coefficient de validité obtenu par une seule étude locale n'est qu'un indicateur imparfait de la validité réelle d'un instrument de mesure. Et même l'accumulation des estimations de la validité à l'aide de plusieurs études ne permettra pas de cerner avec certitude la part d'erreur et de validité réelle. C'est pourquoi il faut également consulter les résultats des méta-analyses appropriées. Plusieurs références relativement à des

17. Lors d'un litige concernant l'application de la Loi sur l'emploi dans la fonction publique (article 21) et mettant en cause Revenu Canada, le Président du Comité d'appel a accepté l'argument de la défense suivant lequel la note de passage n'était pas suffisamment justifiée et a ordonné au ministère de procéder à une étude locale de validation critériée (Lecours, TAX-0583, 1993).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

méta-analyses importantes sur différents instruments de sélection ont déjà été citées dans ce chapitre (voir section « Consulter les résultats des méta-analyses »).

AUTRES MÉTHODES DE VALIDATION

Il est souvent impossible pour le praticien de recourir à l'une ou à l'autre des démarches de validation. En effet, effectuer une étude locale exige un échantillon de taille adéquate, ce qui est impraticable dans bien des organisations de taille modeste ou ne comportant pas suffisamment de personnes par catégorie d'emplois. De la même manière, la généralisation de la validité est parfois impossible, simplement parce qu'il n'existe pas de méta-analyses pour tous les postes et pour tous les prédicteurs, ou insuffisante, parce que les méta-analyses existantes ne comportent pas toutes les informations pertinentes pour satisfaire aux exigences légales.

Validation par les composantes de l'emploi. Hoffman et McPhail (1998) rappellent qu'il existe d'autres démarches pour estimer la validité reliée au critère. L'une d'entre elles consiste, sous certaines conditions stipulées par la Equal Employment Opportunity Commission (1978), à « transporter » les résultats obtenus dans une étude locale de validation provenant d'une situation comparable, si toutefois une telle étude existe. Il va sans dire que cette approche équivaut, au mieux, à faire une étude locale, qui elle-même ne permet pas des estimations de validité très fiables. Hoffman et McPhail proposent plutôt de recourir à une autre méthode, appelée « validité des composantes de l'emploi » (*job component validity*). Dérivée de l'approche de la validité synthétique, cette méthode repose sur le raisonnement suivant : 1) lorsque des emplois ont une tâche commune (p. ex., dresser un budget ou superviser du personnel de production), les qualités requises pour accomplir cette tâche sont semblables pour chacun de ces emplois ; 2) par conséquent, un instrument mesurant l'une de ces qualités sera valide pour l'ensemble de ces emplois. Cette approche requiert l'utilisation du *Position Analysis Questionnaire* (McCormick, Mecham et Jeanneret, 1989), et n'est appliquée pour l'instant que pour les aptitudes mesurées par la *Batterie générale de tests d'aptitudes* (BGTA).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Validation en combinant deux types de preuve. Enfin, si aucune démarche empirique n'est possible et qu'il n'existe pas de méta-analyses pertinentes, Guion (1998) propose une démarche pour étayer la relation prédicteur-critère ; cette démarche combine une preuve basée sur un raisonnement logique et une autre basée sur la validation psychométrique. Premièrement, le prédicteur en cause doit effectivement être jugé comme étant une caractéristique essentielle à la réussite d'aspects importants de l'emploi. Deuxièmement, la validité de l'instrument à mesurer ce prédicteur doit être démontrée. Si la logique et des données soutiennent ces deux affirmations, alors les scores à cet instrument de mesure constituent un prédicteur valide de ces aspects de l'emploi. On retrouve cette démarche de validation dans les *Standards* (1999, p. 155-157)[18].

JUSTIFICATION DES INSTRUMENTS DE SÉLECTION

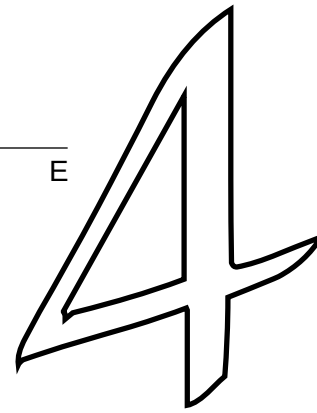
Le processus de validation terminé, il reste à décider si l'instrument de mesure ainsi étudié doit être utilisé (voir étape 5, figure 3.1). La validité est certes une qualité importante : un instrument qui n'est pas valide ne peut pas être utile. Cependant, la validité ne dit pas tout, il y a d'autres enjeux à considérer. L'appréciation de tout instrument de mesure doit passer par une analyse complète de ses avantages et de ses inconvénients et, si possible, par une évaluation de sa rentabilité (voir chapitre 1).

18. Cette démarche correspond au cheminement combinant les inférences 2 et 3 (voir *Standards*, 1999, figure, p. 153).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

C H A P I T R E



FIDÉLITÉ ET CONTRÔLE DES ERREURS DE MESURE

Pour être utile, une mesure doit d'abord être fiable, exacte et digne de confiance. Lorsqu'un patient apprend qu'il a une pression artérielle élevée ou un taux de cholestérol excédant les normes acceptables, il veut une certaine assurance que ses résultats reflètent bien son état de santé. En cas d'erreur de procédure de la part de son médecin ou d'une méprise au laboratoire, il serait superflu, voire dangereux pour lui, de s'astreindre au traitement recommandé. Il en est de même en gestion des ressources humaines. Pour choisir les meilleurs candidats à un poste, par exemple, les résultats aux divers instruments de sélection ne doivent pas être modifiés par la chance, l'humeur du comité de sélection, la fatigue du correcteur ou tout autre facteur étranger aux qualités requises chez les candidats.

Le concept de fidélité concerne le degré d'exactitude des mesures ; on s'y réfère pour désigner la précision, la fiabilité, la constance des résultats ou des scores obtenus à un instrument de

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

mesure. Une mesure fidèle est une mesure sans erreur. Mais qu'entend-on exactement par erreur ? De quel type d'erreur s'agit-il ? Il y a des erreurs fortuites, accidentelles comme une maladresse, un trou de mémoire passager, une distraction, une question ambiguë, certains évaluateurs plus sévères. Il y a aussi des erreurs plus systématiques, comme une qualité qui n'aurait pas dû être mesurée (p. ex., des questions qui exigent des connaissances en géographie dans un examen de mathématiques), un patron raciste ou un candidat toujours anxieux peu importe l'examen.

Plan du chapitre. Connaître les diverses sources d'erreur et la manière dont elles affectent la précision et la stabilité d'une mesure constitue l'objet de la première section de ce chapitre, la plus importante pour le praticien. Non seulement cette connaissance permet d'éclairer la notion de fidélité, mais surtout elle renseigne sur les moyens à prendre pour contrôler et éviter ces erreurs si dommageables en gestion des ressources humaines. Les autres parties traitent de la définition de la fidélité, des méthodes pour l'estimer, des usages pratiques du coefficient de fidélité et des paramètres servant à en apprécier l'ampleur.

SOURCES D'ERREURS DE MESURE ALÉATOIRES ET MOYENS POUR LES CONTRÔLER

L'erreur fait partie de la vie, particulièrement lorsqu'il s'agit de mesurer. Deux personnes qui mesurent un objet avec le même ruban à mesurer peuvent trouver des résultats légèrement différents. Lors d'une élection, personne ne s'étonne d'une différence de quelques votes à la suite d'un recomptage dans une circonscription. Les caractéristiques psychologiques et les dimensions du comportement humain sont encore plus propices à l'erreur. Des concepts aussi complexes et abstraits que la motivation ou le leadership sont pourtant l'objet de tests, de simulations ou d'entrevues ; des connaissances de toutes sortes sont fréquemment évaluées par des examens pratiques ou théoriques ; diverses facettes du rendement au travail sont appréciées tous les jours dans nos organisations lors de processus d'évaluation du rendement. Dans la plupart de ces situations, les résultats obtenus traduisent probablement en bonne partie les variables mesurées, mais, dans tous les cas, une portion d'erreur de mesure, grande ou petite, modifie ces résultats.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Les causes de cette erreur sont multiples. En effet, de nombreux facteurs peuvent altérer la précision ou la stabilité des mesures utilisées en gestion des ressources humaines. Pour nous aider à découvrir ces diverses sources d'erreurs et surtout pour apprendre à éviter ces erreurs dans les situations pratiques les plus courantes, nous allons partir d'un exemple. Supposons qu'une organisation a conçu deux examens pour évaluer exactement les mêmes connaissances chez les candidats. Ces deux examens, sans être identiques, comportent des questions équivalentes au regard de la difficulté et du contenu sur les connaissances à mesurer. D'un concours à l'autre, l'organisation peut ainsi alterner les versions d'examen afin de restreindre la divulgation des questions par les répondants aux candidats suivants. De plus, si elle doit évaluer de nouveau un candidat, l'organisation peut utiliser l'autre version d'examen. Imaginons maintenant que ces deux examens équivalents (ou parallèles) sont administrés aux mêmes personnes ; on devrait alors s'attendre à une corrélation parfaite entre les résultats obtenus à chacune des versions de l'examen, à condition bien entendu que leur niveau réel de connaissances n'ait pas changé entre-temps. Chaque personne obtiendrait exactement le même résultat aux deux versions de l'examen[1]. Toute différence entre les résultats serait due à ce que l'on appelle communément des erreurs de mesure (Magnusson, 1966 ; *Standards*, 1999). Ces erreurs étant de nature aléatoire (Kerlinger, 1986), il est plus juste de les appeler « erreurs de mesure aléatoires » (voir Nunnally et Bernstein, 1994).

Le tableau 4.1 rassemble les principales sources d'erreurs aléatoires de mesure en gestion des ressources humaines. C'est une synthèse de plusieurs auteurs (Cronbach, 1990 ; Dunnette, 1966 ; Gatewood et Feild, 1998 ; Guion, 1998 ; Magnusson, 1966 ; Nunnally et Bernstein, 1994 ; Selltiz, Wrightsman et Cook, 1976). Le tableau présente aussi des moyens pour contrôler les sources d'erreurs ou pour en atténuer les effets. Voyons cela plus en détail.

1. En fait, une corrélation parfaite ne signifie pas que les résultats soient strictement identiques ; l'ajout ou le retrait d'une constante serait sans effet. Par exemple, si tous les candidats obtiennent, disons, cinq points de plus lors du deuxième examen, la corrélation demeurerait parfaite, parce que la position relative de chaque candidat par rapport au groupe serait inchangée ; le candidat qui avait par exemple 10 points de plus que la moyenne conserverait cette avance.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 4.1

PRINCIPALES SOURCES D'ERREURS DE MESURE CONTRIBUANT À L'INSTABILITÉ ET À L'IMPRÉCISION DES RÉSULTATS OBTENUS À DES INSTRUMENTS DE MESURE ET MOYENS POUR LES CONTRÔLER

SOURCES D'ERREURS	MOYENS DE CONTRÔLE
A) Le candidat	
1. Tendance à répondre au hasard	
a) Les réponses **à vue de nez**, les tentatives pour deviner (*guessing*) lorsque le candidat ne comprend pas la question ou ne connaît pas la réponse introduisent un effet de hasard.	– Correction négative. – Évaluation de la constance. – Augmentation du nombre d'items.
b) Les réponses **par pur hasard**, pour se débarrasser, donnent des résultats aléatoires.	– Motivation du répondant par divers moyens. – Évaluation de la constance. – Augmentation du nombre d'items.
2. État général mais passager du candidat	
a) **L'humeur** du moment, l'état de **fatigue** ou de **santé** du candidat sont des facteurs qui peuvent nuire à sa performance.	– Possibilité de remise en cas d'indisposition.
b) La **motivation** du candidat à faire de son mieux ou, au contraire, son manque d'intérêt à répondre aux questions peut avoir un effet sur sa performance.	
c) La **réaction au stress** n'est ni la même chez tous les candidats ni constante chez un même candidat ; selon le stress ressenti, la performance peut être différente.	– Période d'adaptation et ambiance détendue.
d) L'état de **préparation mentale**, de concentration, agit sur le degré d'attention, sur la compréhension des questions ou des problèmes à résoudre et, par conséquent, sur la performance.	

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Note : Il ne faut pas confondre les effets de ces états passagers avec ceux des changements réels dus à la nature dynamique de la caractéristique ou de la dimension à mesurer, telle l'amélioration de la performance par l'apprentissage ou une fluctuation véritable de l'opinion mesurée.

3. Réactions spécifiques fortuites ou inhabituelles du candidat

a) La **compréhension des directives** peut venir plus aisément à certains moments ou chez certaines personnes. – Directives claires.

b) Les **trous de mémoire**, les oublis momentanés, font en sorte qu'une personne manque des questions auxquelles elle aurait pu répondre correctement. – Augmentation du nombre d'items.

c) Les **fautes d'inattention,** les distractions lors de la lecture ou de l'écoute des questions, au moment de formuler ou d'inscrire la réponse, nuisent aux résultats. – Augmentation du nombre d'items.

4. Interaction avec les caractéristiques de l'examinateur

Il arrive que des candidats réagissent différemment en fonction de certaines caractéristiques de l'examinateur (le **sexe**, la **race**, l'**âge**, un trait de **personnalité**, etc.) ; cet effet d'interaction peut être important lors d'un instrument interactif appliqué individuellement (p. ex., en entrevue). – Choix d'un examinateur « neutre ».

B) L'examinateur

Les sources d'erreurs appartenant à cette catégorie ont une influence plus grande lorsque l'application de l'instrument de mesure exige une participation soutenue de l'examinateur (p. ex., en entrevue) ou lorsque le processus d'évaluation requiert son jugement ou son attention (p. ex., questions ouvertes, observation lors d'une simulation).

5. État général mais passager de l'examinateur

L'examinateur n'est pas à l'abri des fluctuations passagères de son état général ; son **humeur**, son niveau de **fatigue**, sa condition de **santé**, sa **motivation** ou sa **préparation mentale** peuvent varier ; son comportement et celui du candidat en seront peut-être affectés.
– Possibilité de remise en cas d'indisposition.
– Standardisation.
– Augmentation du nombre d'examinateurs.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 4.1 *(suite)*

PRINCIPALES SOURCES D'ERREURS DE MESURE CONTRIBUANT À L'INSTABILITÉ ET À L'IMPRÉCISION DES RÉSULTATS OBTENUS À DES INSTRUMENTS DE MESURE ET MOYENS POUR LES CONTRÔLER

SOURCES D'ERREURS	MOYENS DE CONTRÔLE
6. Réactions spécifiques fortuites ou inhabituelles de l'examinateur	
Les réactions fortuites, comme les **fautes d'inattention** (p. ex., lors de la lecture ou de l'écoute des réponses, au moment de prendre des notes, d'inscrire une cote ou simplement d'en faire la compilation) ou une **difficulté momentanée à comprendre** certains propos du répondant affectent les résultats.	– Augmentation du nombre d'examinateurs.
7. Interaction avec les caractéristiques du répondant	
En fonction des caractéristiques des candidats (le **sexe**, la **race**, l'**âge**, les traits de **personnalité**, le lien de **parenté**, etc.), l'examinateur peut agir différemment ou manquer d'objectivité et d'équité dans le traitement de leurs réponses.	– Augmentation du nombre d'examinateurs. – Sensibilisation aux biais de perception.
C) La situation	
8. Conditions d'administration non standardisées	
La **durée**, le **matériel** disponible, **l'ambiance**, la **température**, l'**éclairage**, le **bruit**, les **distractions** et autres conditions qui entourent l'utilisation de l'instrument de mesure peuvent influencer la performance du candidat ; des conditions changeantes ou non conformes à ce qui a été prévu peuvent altérer la performance.	– Standardisation des conditions.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

D) L'instrument

9. Composantes non standardisées	
a) La composition de l'instrument de mesure détermine fortement les réponses formulées par les candidats et les résultats obtenus; c'est pourquoi des **items** changeants ou non conformes à ce qui a été prévu (contenu des questions, formulation des problèmes, choix de réponse proposés, agencement de ces éléments, etc.) ou des **directives** non standardisées constituent une source de variation indésirable dans la performance des candidats.	– Standardisation des items et des directives.
b) En ce qui concerne les **outils d'évaluation** ou leur application, une standardisation insuffisante peut engendrer de l'instabilité dans les résultats.	– Standardisation des outils d'évaluation.
10. Ambiguïtés et insuffisances	
a) Si une formulation est ambiguë, trop complexe ou incomplète dans un **item** ou dans les **directives**, les candidats peuvent ne pas faire tous, ni toujours, la même interprétation de ce qui leur est demandé; ces fluctuations, bien qu'erratiques, ont une incidence sur leurs réponses.	– Items et directives clairs et complets. – Questions d'éclaircissement. – Entraînement à l'aide de questions types.
b) Si les **outils d'évaluation** sont ambigus ou incomplets, les examinateurs sont laissés à eux-mêmes, sans garantie qu'ils auront tous et toujours une manière analogue de traiter les réponses des candidats; une certaine variation inéquitable dans les résultats est alors possible.	– Outils d'évaluation clairs et complets. – Augmentation du nombre d'examinateurs. – Formation des examinateurs.
11. Échantillonnage des items (*content sampling*)	
Il arrive qu'un candidat réussisse mieux certains items que d'autres. Par ailleurs, les items d'un instrument de mesure ne sont qu'un **échantillon parmi tous les items possibles** du domaine à mesurer (p. ex., il y a une infinité de problèmes d'addition, d'orthographe, de comptabilité, de négociation, etc.). Or, si un échantillon différent d'items avait été retenu pour composer l'instrument, les résultats obtenus auraient pu être différents également.	– Augmentation du nombre d'items.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

LE CANDIDAT

Une première source d'erreurs de mesure aléatoires est le candidat lui-même, qui n'est pas toujours des plus constants dans sa façon de répondre à un instrument de mesure ; cela entraîne imprécision et instabilité dans ses résultats.

1. Tendance à répondre au hasard. Un candidat qui ne comprend pas une question qui lui est posée ou qui ne connaît pas la réponse peut décider de tenter sa chance et répondre « **à vue de nez** ». Si des choix de réponses lui sont fournis, il peut essayer de deviner laquelle est la meilleure ; sinon, il répondra de son mieux. Parfois il est chanceux et tombe sur la bonne réponse, parfois il ne l'est pas. Si un candidat répond ainsi aux deux versions équivalentes de l'examen de notre exemple, il obtiendra des résultats différents suivant sa chance à chaque passation. Il peut arriver également qu'un candidat donne des réponses complètement **au hasard**, sans même tenter de deviner la réponse. Cela peut se produire lorsqu'un candidat n'est pas particulièrement motivé ni intéressé par l'instrument de mesure, comme dans le cas d'un inventaire de personnalité qui comporterait de nombreuses questions personnelles ou d'une entrevue de sélection dirigée par un comité paraissant mal organisé. On relève aussi ce manque de motivation chez les superviseurs qui doivent évaluer leur personnel avec des outils qu'ils jugent inutilement lourds et compliqués. Ces façons de répondre, « à vue de nez » ou simplement au hasard, vont évidemment produire des différences dans les résultats obtenus d'un examen à l'autre, contribuant ainsi à l'erreur de mesure.

Il n'y a pas vraiment de moyens pour éliminer les réponses « à vue de nez », à moins d'aviser le répondant que des points seront enlevés pour toute réponse erronée. Cependant, la tendance à répondre au hasard, pour se débarrasser, peut être réduite notamment en se souciant de la motivation du répondant à se conformer à ce qui lui est demandé. Ce dernier doit pouvoir saisir la pertinence du contenu de l'instrument par rapport aux objectifs poursuivis (soit la validité apparente) ; il ne doit pas y voir d'éléments jugés inutiles ni de difficultés indues. En outre, l'examinateur ou l'organisation qu'il représente doit inspirer confiance. Dans certains inventaires de personnalité, on retrouve des énoncés servant à évaluer la constance du répondant.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Premier principe général pour contrôler les erreurs aléatoires de mesure : l'augmentation du nombre de mesures. S'il n'est pas facile de contrecarrer ces tendances à répondre au hasard, il est relativement simple d'en atténuer les effets sur les résultats obtenus, en augmentant le nombre de mesures effectuées. Pensons, par exemple, à un examen de connaissances pour lequel il n'y aurait qu'une seule question « vrai ou faux ». Si le répondant ne connaît pas la réponse, son score réel, sans erreur, devrait être zéro. Mais, en raison des probabilités, ce répondant a tout de même une chance sur deux d'obtenir la bonne réponse par pur hasard. On sait cependant qu'il ne peut pas être toujours aussi chanceux. Si l'on augmente le nombre d'énoncés à deux, ses probabilités passent à une chance sur quatre d'obtenir un score parfait ($1/2 \times 1/2$) ou à une chance sur huit ($1/2 \times 1/2 \times 1/2$) avec trois énoncés. Ce principe s'applique aussi aux choix de réponses. Si l'on change les questions « vrai ou faux » par des questions à choix multiples comprenant quatre choix de réponses, alors le candidat n'a plus qu'une chance sur quatre d'obtenir la bonne réponse par pur hasard à chaque question ; l'effet du hasard se trouve réduit de moitié en doublant le nombre de choix de réponse.

Ainsi, *augmenter le nombre d'items diminue les probabilités d'obtenir un résultat éloigné de la réalité et, réciproquement, augmente celles de s'en rapprocher.* En effet, il est peu probable, statistiquement parlant, d'être systématiquement chanceux ou systématiquement malchanceux, tantôt on est chanceux et on obtient par hasard une réponse que l'on ne connaissait pas. Tantôt c'est l'inverse et on se trompe pour une réponse que l'on savait pourtant. Lorsqu'elles s'additionnent, les erreurs dues au hasard finissent par se neutraliser d'elles-mêmes avec la répétition. En effet, comme elles suivent une distribution normale parfaitement symétrique, la somme de ces erreurs tend vers zéro (voir le chapitre 7). Bref, plus on a de mesures équivalentes pour une même variable, moins la somme de ces mesures sera affectée par les erreurs aléatoires et plus elle se rapprochera de la réalité (ce principe est abordé plus loin, à la section portant sur les facteurs pouvant influencer le coefficient de fidélité, sous la rubrique « Nombre d'énoncés et longueur de l'instrument »).

Il est intéressant de remarquer que l'ajout de mesures comme moyen de contrôle agit sur les effets de la source d'erreurs, mais pas sur la source elle-même. En effet, augmenter le nombre d'items n'élimine pas le hasard et les erreurs qui lui sont imputables au regard

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

de chaque mesure prise isolément. La probabilité sera toujours de 50 % d'obtenir, par hasard, la bonne réponse à une question « vrai ou faux ». Toutefois, l'addition de plusieurs de ces mesures permet aux erreurs de s'annuler les unes les autres et finit par en neutraliser les effets.

Ce principe est très important en ce qui concerne le contrôle des erreurs aléatoires et il est d'application simple pour à peu près tous les instruments de mesure. Ainsi, il vaut mieux : 1) ne pas se limiter à une ou deux questions d'entrevue ou d'examen par dimension évaluée ; 2) dans une entrevue comportementale, demander plus d'un exemple par question ; 3) dans une entrevue situationnelle, soumettre plus d'une mise en situation par dimension évaluée ; 4) procéder à deux ou même trois entrevues plutôt qu'une (au moins pour les quelques candidats finalistes) afin de réduire la possibilité d'une performance non représentative de la part des candidats ; 5) augmenter le nombre de membres participant à un comité de sélection pour diminuer les erreurs aléatoires de perception et d'interprétation ; 6) faire corriger à l'aveugle par deux examinateurs différents un même examen ; 7) vérifier les références auprès de plus d'un ancien employeur ou même de plusieurs personnes par employeur ; 8) lors d'une évaluation du rendement, multiplier les sources d'informations, etc. Évidemment, augmenter le nombre de mesures comporte un coût en termes de temps et d'énergie. De plus, nous verrons plus loin que l'amélioration relative apportée par cette technique diminue à mesure que le nombre de mesures augmente (p. ex., ajouter un énoncé à un examen qui en compte déjà 100 n'aura plus le même effet que de passer de un à deux énoncés).

2. État général mais passager du candidat. Plusieurs dispositions ou états, tels que l'**humeur**, la **fatigue**, la **santé**, changent d'un jour à l'autre ou d'un candidat à l'autre, alors qu'ils peuvent influencer leur performance à un instrument de mesure. De pareilles fluctuations s'observent aussi dans la **motivation** du répondant à réussir ou à faire de son mieux, dans sa **réaction au stress** ou dans sa **préparation mentale**. Ces divers états passagers sont connus pour avoir une incidence sur la concentration et les efforts du répondant, faisant fluctuer arbitrairement ses résultats d'une fois à l'autre.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Il existe des moyens pour contrôler ces sources d'erreurs de mesure, certains étant toutefois moins praticables que d'autres en situation de gestion des ressources humaines. Un examinateur pourrait par exemple éviter de donner un examen ou de procéder à une entrevue lorsqu'une personne éprouve des problèmes passagers de santé, de fatigue ou d'ordre émotionnel. Avec tous les inconvénients que cela peut représenter, il pourrait alors y avoir une séance de reprise. Pour permettre au candidat de contrôler son stress en situation d'évaluation, il est courant de commencer par un accueil cordial et chaleureux, de prévoir une période d'adaptation, le temps que le candidat se sente suffisamment à l'aise. Une fois l'évaluation commencée, les sources de stress devraient être minimisées, à moins de vouloir justement mesurer ses effets sur le candidat. On peut même faire une pause ou une diversion en cours d'entrevue pour détendre un candidat trop nerveux.

Toute fluctuation de résultats entre deux situations ne doit pas d'emblée être interprétée comme de l'erreur de mesure : il arrive que des **changements réels** surviennent pour une caractéristique mesurée. Ainsi, il peut y avoir eu apprentissage des connaissances ou des habiletés évaluées, modification de l'attitude ou de l'opinion sondée, changement profond du trait de personnalité ou du style de gestion mesuré. Réciproquement, l'absence de fluctuation ne signifie pas nécessairement que la mesure est exempte d'erreur. C'est le cas de la **mémorisation** de certaines questions et réponses lors de la première passation de l'instrument, qui a pour effet d'augmenter artificiellement la constance avec la deuxième passation.

3. Réactions spécifiques fortuites ou inhabituelles du candidat. D'autres inconstances chez le candidat peuvent perturber ses résultats. Ce sont des réactions particulières à certains aspects ou à certains items de l'instrument de mesure. Parfois, un candidat éprouve de la difficulté à **comprendre les directives** ; il bute sur un mot, une expression ou se fourvoie pour une raison inconnue. Personne non plus n'est à l'abri d'un **trou de mémoire** inopportun, empêchant momentanément de donner une réponse pourtant sue. Dans un moment de distraction, il arrive aussi que des **fautes d'inattention** soient commises : on se méprend sur le sens d'une question ou on donne une réponse alors que c'est une autre qu'on a à l'esprit. Ces réactions fortuites introduisent un effet de hasard dans les résultats

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

d'un candidat et contribuent à l'erreur de mesure. De plus, si leurs fréquences fluctuent d'un examen à l'autre, des différences de résultats seront observées.

La façon de s'y prendre pour atténuer ces erreurs aléatoires de mesure, du moins pour les trous de mémoire et les fautes d'inattention, a déjà été présentée : il s'agit d'augmenter le nombre d'items pour diminuer l'effet du hasard. Pour les difficultés ponctuelles de compréhension des directives, il n'y a pas grand-chose à faire, à part recourir à des directives claires et standard pour tous et en toutes situations.

4. Interaction avec les caractéristiques de l'examinateur. Chaque individu réagit différemment devant une autre personne ; il en est de même du candidat devant l'examinateur. Selon le **sexe**, la **race**, l'**âge** ou le type de **personnalité** de l'examinateur, certains candidats se sentiront plus en confiance, plus disposés à faire de leur mieux, alors que d'autres, devant le même examinateur, éprouveront de la nervosité ou seront intimidés. Comme chaque candidat est différent, il est impossible d'agir sur cet effet d'interaction. On pourra au moins tenter de faire appel à un examinateur dont, semble-t-il, les caractéristiques personnelles risquent peu de déranger un grand nombre de candidats.

L'EXAMINATEUR

L'examinateur peut être une source importante d'imprécision dans les résultats, surtout lorsque l'application de l'instrument de mesure exige sa participation soutenue ou lorsque le processus d'évaluation requiert son jugement ou son attention. Par exemple, lors d'un test individuel, d'une entrevue ou de toute autre forme d'évaluation mettant en interaction étroite le candidat et l'examinateur, les **comportements** de ce dernier (l'accueil, la manière de mettre le candidat à l'aise, sa façon de poser les questions ou ses réactions aux réponses) peuvent avoir un effet sur la performance de certains candidats. Ou encore, pour de nombreux instruments de mesure, l'**évaluation des réponses** du candidat est soumise, à des degrés variables, au discernement et au jugement de l'examinateur. C'est le cas de l'entrevue de sélection, incontestablement l'outil le plus répandu en gestion des ressources humaines. Mais n'oublions pas les examens écrits comportant des questions ouvertes, les cas à résoudre et les simulations dont le but

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

est d'observer le candidat en action. Dans toutes ces situations, la performance du candidat et ses résultats peuvent être affectés par la constance de l'examinateur, dont le comportement est aussi influencé par plusieurs facteurs. Ces facteurs, analogues à ceux présentés pour le répondant, concernent l'état général mais passager de l'examinateur, ses réactions fortuites et une possible interaction avec les caractéristiques du répondant.

5. État général mais passager de l'examinateur. Comme tout le monde, l'examinateur est sujet à des fluctuations passagères de son état général : son **humeur**, son degré de **fatigue**, sa condition de **santé**, sa **motivation** ou sa **préparation mentale** peuvent varier. Un examinateur préoccupé par un souci personnel, épuisé par de longues heures de travail ou qui n'a pas pris le temps de se concentrer risque de ne pas porter suffisamment attention au répondant, de ne pas être à l'écoute ou d'avoir plus de difficulté à réagir adéquatement à ce qui se passe. En outre, lors du processus d'évaluation, son jugement pourrait être altéré.

Afin de contrôler cette source potentielle d'erreur, diverses dispositions peuvent être prises. L'examinateur peut éviter de rencontrer des candidats ou de corriger si son état affecte ses facultés. Il peut se conditionner à l'importance de procéder à ces évaluations. Il peut standardiser certains aspects qui influent sur son état, comme par exemple faire ses entrevues le matin, alors qu'il est en meilleure forme, ou planifier un horaire réaliste, à la mesure de son énergie. La standardisation est un moyen important dans le contrôle de l'erreur de mesure ; elle est présentée plus en détail à la section suivante portant sur la situation.

Il existe un autre moyen, sans doute le plus efficace, pour contrôler les effets néfastes de ces fluctuations passagères de l'état de l'examinateur. Basé sur le principe de l'augmentation du nombre de mesures, ce moyen consiste à recourir à plus d'un examinateur. Par exemple, un comité de sélection, en comparaison à un interviewer unique, a moins de chances que tous ses membres soient victimes en même temps d'un malaise passager ou aient de la difficulté à se concentrer. Procéder à une double correction à l'aveugle par deux examinateurs est également efficace pour neutraliser une partie des inconstances ; cette méthode est d'ailleurs appliquée dans le processus de correction pour un des examens d'entrée à une corporation

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

professionnelle fort connue au Canada. Une fois les deux corrections effectuées, si l'écart entre les deux résultats pour un même candidat ne dépasse pas un certain seuil, le score attribué à ce candidat est la moyenne des deux résultats ; si l'écart excède le seuil prévu, la correction est alors reprise par un troisième examinateur.

6. Réactions spécifiques fortuites ou inhabituelles de l'examinateur. L'examinateur peut avoir des réactions fortuites ; un **manque d'attention** ou une **difficulté à comprendre** certains propos du répondant est toujours possible. Par inadvertance, il peut parfois commettre une erreur au moment de prendre des notes, d'inscrire une cote ou simplement d'en faire la compilation. Ces incidents ont évidemment pour effet d'introduire de l'erreur dans les résultats. Étant donné leur nature imprévisible, ils sont difficiles à éliminer ou à contrôler. Cependant, comme nous l'avons démontré au paragraphe précédent, il est possible d'en neutraliser les effets de façon assez efficace en augmentant le nombre d'examinateurs. Plus ils sont nombreux lors de l'application de l'instrument ou du processus d'évaluation, moins il y a de chances qu'ils aient en même temps les mêmes réactions néfastes.

7. Interaction avec les caractéristiques du répondant. Enfin, il arrive également que le comportement de l'examinateur varie en fonction de certaines caractéristiques du candidat, comme par exemple le **sexe**, la **race**, l'**âge**, certains traits de **personnalité** ou le lien de **parenté** avec lui ou avec quelque autre personne significative. Par exemple, dans le cadre d'une entrevue, il se peut que l'examinateur ait tendance à être plus chaleureux avec un candidat de sexe opposé, moins inquisiteur pour quelqu'un de sa race, plus sur la défensive avec un candidat particulièrement jeune ou qu'il se prenne de sympathie pour une personne handicapée. L'examinateur peut aussi faire preuve d'un manque d'objectivité lors de l'évaluation, de façon consciente ou non, et se laisser influencer par des caractéristiques du candidat qui n'ont rien à voir avec l'objet mesuré par l'instrument.

Encore une fois, le meilleur moyen de se prémunir contre cette source d'erreur est le recours à plusieurs examinateurs. On peut aussi miser sur la sensibilisation des examinateurs aux différentes sources de biais de perception et autres stéréotypes, sans toutefois avoir de garantie quant à l'efficacité réelle de cette mesure.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

LA SITUATION

Les conditions dans lesquelles est appliqué un instrument de mesure peuvent avoir une influence sur les résultats des candidats ; si elles ne sont pas maintenues constantes, elles amplifieront l'erreur de mesure.

8. Conditions d'administration non standardisées. L'administration d'un instrument de mesure comporte diverses conditions, comme la **durée** allouée au candidat pour répondre à l'instrument de mesure et, dans plusieurs cas, le **matériel** nécessaire (cahier de test ou d'examen, crayons, papier, tableau, projecteur, outils, pièce d'équipement, véhicule, etc.). Chacun de ces éléments, lorsqu'il varie d'une situation à l'autre ou d'un candidat à l'autre, peut être source d'instabilité dans les résultats. C'est aussi le cas des caractéristiques de l'environnement physique, comme la **température**, l'**éclairage** ou le niveau de **bruit**, sans compter les **distractions** fortuites, particulièrement dérangeantes pour le répondant, qui peuvent être nombreuses : conversations environnantes, passage d'une personne du sexe opposé, sonnerie de téléphone, etc.

Deuxième principe général pour atténuer les erreurs aléatoires de mesure : la standardisation. *Pour éviter que des conditions changeantes deviennent source d'instabilité, il suffit qu'elles soient scrupuleusement contrôlées et maintenues constantes pour tous les candidats à toutes les applications.* C'est le principe de la standardisation qui a pour effet de contrôler directement l'erreur à sa source. Chaque facteur de la situation ayant un effet potentiellement indésirable sur les résultats des répondants devrait être standardisé. Il peut être fort utile, à cet égard, surtout si plusieurs examinateurs ou plusieurs sessions sont en cause, de constituer un manuel de procédures où des balises sont clairement établies notamment en ce qui concerne la durée, le matériel et les autres conditions. Il est également recommandé de procéder dans un local fermé, à l'abri des interruptions et des indiscrétions ; un rappel aux candidats de fermer leur téléphone cellulaire évite bien des distractions.

Il y a des contextes où le principe de la standardisation n'est pas si simple à appliquer. C'est le cas, par exemple, des **mises en situation collectives** qui servent d'ordinaire à évaluer les caractéristiques relationnelles des candidats. Compte tenu de la nature interactive de ces simulations, ce que disent et font les autres participants fait partie de la situation à laquelle fait face un candidat. Or, les paroles et les gestes

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

de ses partenaires varient forcément d'un groupe à un autre. Le climat et le ton de la discussion peuvent être différents entre les groupes de candidats, parfois animés et provocateurs, parfois calmes et ordonnés, selon la composition du groupe et son état d'esprit. Ce manque potentiel de standardisation d'un groupe à l'autre signifie que les évaluateurs ont parfois de la peine à déterminer si les comportements observés chez un candidat sont typiques de cette personne ou fonction de la dynamique du groupe (Thornton, 1992). Cette limitation importante de l'évaluation collective peut être amoindrie, notamment de deux manières. D'une part, on peut réduire le plus possible le nombre de groupes différents en restreignant ce mode d'évaluation à quelques candidats ; du point de vue de la standardisation, avoir un seul groupe serait idéal. D'autre part, il faut avoir un certain nombre de personnes par groupe afin d'augmenter la probabilité d'y retrouver un échantillon le moindrement représentatif et, partant, plus comparable d'un groupe à l'autre. Une chose est certaine : il ne faut pas suivre l'exemple de cette organisation qui avait choisi d'évaluer les candidats par groupe de deux, amenant ainsi chaque candidat à être évalué dans des conditions différentes !

L'INSTRUMENT DE MESURE

Finalement, l'instrument de mesure lui-même peut être la cause d'instabilité et d'imprécision dans les résultats. Ce problème survient principalement lorsque l'instrument est insuffisamment standardisé ou lorsqu'il comporte des ambiguïtés ou des insuffisances. L'échantillonnage des items qui composent le contenu de l'instrument peut aussi être une source d'erreurs de mesure.

9. Composantes non standardisées. Un instrument de mesure est composé d'items (questions, problèmes, mises en situation, etc.), de directives et d'outils d'évaluation. Les **items** et les **directives** adressés au répondant ont pour but de recueillir auprès de ce dernier des réponses (ou des comportements) typiques de son niveau de performance. Ces composantes ont évidemment une incidence déterminante sur les réponses formulées par le répondant. C'est pourquoi des items changeants ou non conformes à ce qui a été prévu (contenu des questions, formulation des problèmes, choix de réponse proposés,

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

agencement de ces éléments, etc.) ou des directives affublées des mêmes maux constituent une source de variation indésirable dans la performance des candidats.

Quant aux **outils d'évaluation**, ils servent à apprécier les réponses recueillies auprès des candidats. Ces outils peuvent comprendre une clé de correction accompagnée de directives pour l'examinateur, d'une procédure de compilation des résultats, d'une grille d'observation ou d'interprétation, etc. Les outils d'évaluation et leur application par l'examinateur ont donc un effet direct sur les résultats obtenus par les candidats. S'ils ne sont pas standardisés, si des inconstances sont observées d'un candidat à l'autre ou d'une occasion à l'autre, il peut alors en résulter des variations fortuites dans les résultats, sans rapport avec le véritable objet mesuré par l'instrument en cause.

Le remède pour éviter pareilles fluctuations est encore celui de la standardisation : plus grande est la standardisation des composantes de l'instrument, moins il y a d'erreurs de mesure aléatoires liées à l'instrument. Par exemple, la clé de correction doit être identique pour tous et en tout temps, afin que chaque candidat et chaque réponse fournie soient évalués uniformément, avec les mêmes éléments de réponse attendus et en fonction des mêmes barèmes ; un code de procédures peut contribuer à cette uniformité d'application. Les directives ne doivent pas changer d'une occasion à l'autre ; si des questions sont posées par des candidats, l'examinateur doit y répondre toujours de la même manière et les éléments eux-mêmes et leur agencement devraient être soumis à la standardisation.

Les instruments les plus standardisés sont naturellement les tests ou les examens écrits, composés de questions fermées et assortis d'une clé de correction unique ; c'est une de leurs qualités distinctives. Il en va autrement d'une entrevue ou de tout autre instrument interactif (p. ex., simulation d'une présentation orale devant un comité dont les membres posent des questions). La souplesse qui est introduite dans ces instruments et certains avantages qui en découlent (p. ex., possibilité d'obtenir une information additionnelle pertinente) le sont au prix d'une diminution de la standardisation et des inconvénients qu'elle comporte. Il revient alors au responsable de créer le difficile équilibre entre standardisation et souplesse, puis d'en évaluer les conséquences par rapport aux objectifs poursuivis.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

10. Ambiguïtés et insuffisances. Il ne sert à rien de standardiser scrupuleusement les composantes d'un instrument si ces dernières sont ambiguës ou incomplètes. Malgré l'attention portée à la formulation des **items** (questions, problèmes, mise en situation, etc.), il y a toujours un risque d'ambiguïté. Un mot, une expression peut soudainement être interprétée par un répondant d'une manière imprévue. Faute d'avoir été suffisamment explicite, une question peut amener un répondant à s'en remettre à des suppositions spontanées. Par exemple, à une question d'entrevue portant sur la résolution d'un conflit entre deux employés, si les répondants ne sont pas informés des mesures qui ont déjà été prises pour régler ce problème, chacun des répondants est laissé à ses propres présomptions quant au stade où en est rendu le conflit, affectant d'emblée sa réponse.

Les **directives** jouent aussi un rôle capital pour le répondant. S'il y subsiste des ambiguïtés ou si elles sont insuffisantes, les candidats n'auront pas tous ni toujours la même façon d'interpréter ce qu'ils doivent faire ou répondre. Par conséquent, ceux qui ont la même compréhension que celle visée par l'instrument de mesure sont avantagés par rapport aux autres qui n'ont pas eu cette chance. Par exemple, on demandait à des candidats dans un examen de fournir des exemples de situations qu'ils avaient vécues et qui illustraient leurs capacités de gestion. Comme les directives omettaient de préciser le nombre d'exemples à fournir, la quantité d'exemples a varié substantiellement d'un candidat à l'autre. Or, ceux qui ont opté pour la quantité avaient plus de chances de décrocher les points prévus à la grille de correction.

Enfin, il y a les **outils d'évaluation**, qui servent essentiellement à transformer les réponses des candidats en résultats. Des outils ambigus ou incomplets engendrent de l'instabilité, cette fois-ci, chez l'examinateur. Laissé à sa propre interprétation du moment, un examinateur peut dériver dans sa façon de traiter les réponses d'un candidat ou d'une situation à l'autre. Ainsi, deux réponses semblables pourront donner lieu à des pointages différents. L'inconstance peut être pire encore si l'on a recours à plusieurs examinateurs.

Troisième principe général pour atténuer les erreurs de mesure aléatoires : des composantes claires et complètes. Un énoncé ambigu ou incomplet peut prêter à des interprétations diverses, et aléatoires jusqu'à un certain point, créant ainsi de l'erreur de mesure (Kerlinger,

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

1986). La solution pour éliminer ce genre d'erreurs de mesure aléatoires est logiquement fort élémentaire : *employer des composantes claires et complètes, univoques et sans possibilité d'interprétation*. À cet égard, on pourra s'en remettre aux nombreuses règles sur la formulation des questions d'examen et des directives (voir chapitre 6). De plus, il est essentiel que les items et les directives soient testés auprès d'un échantillon représentatif de candidats afin de s'assurer qu'aucune ambiguïté ne subsiste. Lors de l'application de l'instrument, il est de mise d'inviter les candidats à poser des questions d'éclaircissement sur les directives. On peut aussi prévoir une possibilité d'entraînement pour les candidats à l'aide d'un échantillon de questions types. Enfin, la même rigueur doit aussi s'appliquer aux outils d'évaluation.

Selon le type d'instruments de mesure, il peut être extrêmement laborieux, voire parfois impossible, de concevoir des outils d'évaluation sans équivoque, complets et qui ne requerront pas une certaine dose d'interprétation de la part de l'examinateur. Par exemple, élaborer une clé de correction pour une question ouverte ou un cas à résoudre n'est pas une mince affaire, surtout si tous les éléments de réponse valables possibles doivent être prévus. Pour y parvenir, il faut avoir recours à des experts du contenu évalué (*subject matter expert*) et, pour plus de prudence, soumettre l'instrument à un échantillon de candidats. Par la suite, une fois l'instrument en application, il faut modifier par écrit la clé de correction au fur et à mesure que de nouveaux éléments de réponse sont acceptés.

Il n'est pas aisé non plus de produire une grille qui servira à évaluer les réponses et les comportements d'un candidat observé lors d'une **simulation** ou d'une **entrevue**. Il faut dès lors identifier exhaustivement les éléments de réponse ou de comportement attendus, puis les traduire en des termes observables et objectifs de manière à réduire au minimum l'interprétation de l'examinateur. Encore une fois, c'est plus facile à dire qu'à faire, comme en témoignent ces quelques exemples d'éléments de réponse tirés de diverses grilles, chacune réalisée pourtant par des spécialistes reconnus :

a) « Présente les problèmes […] avec crédibilité et impact. »

b) « Découvre les causes principales. »

c) « Arrive à des conclusions solides et prend des décisions sensées. »

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

d) « Analyse les problèmes de façon rationnelle en identifiant les éléments clés d'une situation. »

e) « Dégage des solutions pratiques et viables. »

f) « Ne cherche pas à manipuler ou à contrôler le groupe à des fins personnelles. »

L'évaluation d'un candidat avec chacun de ces éléments sous-tend nécessairement un cadre de référence qui peut varier d'un examinateur à l'autre. Effectivement, une présentation perçue comme « crédible » par un évaluateur ne le sera pas nécessairement par un autre. Évaluer des éléments comme « Découvre les causes principales » ou « Arrive à des conclusions solides et prend des décisions sensées » exige que non seulement les évaluateurs connaissent eux-mêmes ces causes principales et ces décisions sensées, mais aussi qu'ils partagent le même point de vue. Et il en va de même pour les autres exemples, où il est extrêmement difficile d'enrayer la subjectivité, sans au préalable avoir fourni aux évaluateurs un répertoire détaillé des réponses attendues ou souhaitables telles qu'elles auront été déterminées par des experts du domaine.

La subjectivité ne concerne pas que les éléments de réponse attendus. Elle se retrouve également dans l'échelle de cotation mise à la disposition de l'examinateur pour exprimer son évaluation proprement dite des éléments de réponse fournis par un candidat. Comportant généralement de 3 à 10 gradations, ce genre d'échelle exige beaucoup de jugement de la part des examinateurs. Par exemple, pour les éléments *d*, *e* et *f* mentionnés plus haut, l'examinateur devait évaluer si, en fonction des réponses fournies, le candidat méritait la cote « traité », « partiellement traité » ou « non traité ». Il est clair que pour effectuer une telle évaluation, chaque examinateur risque de s'en remettre à ses propres normes pour déterminer, de son mieux, à quel niveau de performance correspond chacune de ces gradations. La formation des examinateurs est l'un des moyens utilisés pour diminuer un tant soit peu la subjectivité des examinateurs et améliorer d'autant leur constance. Le recours à un comité plutôt qu'à un seul examinateur est une autre façon d'y parvenir. En vertu du principe de l'augmentation du nombre de mesures, plusieurs examinateurs permettent ainsi de neutraliser l'effet potentiel des diverses interprétations.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

11. Échantillonnage des items *(content sampling)***.** En plus des ambiguïtés toujours possibles ou d'une standardisation imparfaite, une autre cause d'erreur trouve sa source dans les items : leur échantillonnage. Un instrument de mesure comprend effectivement un nombre limité d'items, par ailleurs représentatifs, en principe, des innombrables items qui appartiennent au domaine de contenu à mesurer. Il y a en effet une infinité de problèmes d'addition, d'orthographe, d'analyse logique, de comptabilité, un nombre incalculable de situations de négociation, de conduite automobile, d'assemblage de pièces de toutes sortes, etc. Conséquemment, de tous les items pouvant théoriquement faire partie d'un instrument de mesure, seulement quelques-uns composent réellement l'instrument. Il faut reconnaître qu'il y a là une certaine dose de hasard : d'autres items auraient pu aussi bien être choisis et donner un instrument légèrement différent.

Or, il arrive qu'un candidat réussisse mieux certains items que d'autres, pour diverses raisons plus ou moins fortuites. Par exemple, un étudiant qui a été particulièrement intéressé par un personnage historique aura de meilleurs résultats à l'examen d'histoire si, par chance, il y a des questions sur ce personnage. Si, lors d'un examen de reprise, l'échantillon de questions choisi par son professeur était différent, les résultats de cet étudiant pourraient en être affectés, même si son niveau de connaissances est demeuré inchangé.

L'effet de hasard lié à l'échantillonnage des items diminue en fonction du nombre d'items : c'est le meilleur moyen de contrôler cette source d'erreur. Réciproquement, moins il y a d'items dans un instrument de mesure, plus les chances de dérapage sont élevées. Un cas extrême est fourni par cette organisation qui a mis sur pied un examen pour mesurer les capacités d'analyse logique ; l'examen était constitué d'une mise en situation générale, suivi de trois questions à développement. Pour répondre à ces questions, il fallait absolument comprendre le mise en situation de départ qui, pour des fins de validité apparente, portait sur un réseau informatique. Les résultats furent étonnants : un candidat a eu une note frisant zéro, alors qu'il occupait déjà avec succès un poste similaire ailleurs. Après analyse, l'organisation s'est rendu compte que la mise en situation de départ exigeait la connaissance d'un élément particulier en informatique ; par ailleurs, cette exigence ne devait pas causer de problème puisqu'on présumait que tous les candidats détenaient cet élément de connaissance. Pas de chance, cet élément était inconnu du candidat,

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

ce qui l'a empêché de répondre correctement aux questions. En réalité, c'est comme si l'examen ne comportait qu'un seul item ; le candidat n'a aucune chance de se reprendre en cas d'échec. La morale de cette histoire est qu'il faut se méfier d'une seule question à développement, aussi générale soit-elle, d'une entrevue unique, d'une seule source de référence ou d'une brève période d'observation.

Résumé. Pour des raisons tout à fait étrangères à ce qui est censé être mesuré, un candidat peut obtenir des résultats différents à deux versions équivalentes du même examen ou au même examen passé à deux occasions distinctes. De telles différences de résultats sont inévitables, même si la caractéristique ou les compétences mesurées chez le candidat n'ont pas changé dans la réalité. Appelées erreurs de mesure, ou plus précisément erreurs de mesure aléatoires, ces variations sont des erreurs de nature accidentelle : les résultats sont affectés parfois dans un sens, parfois dans l'autre. Elles sont occasionnées par les nombreuses sources d'instabilité présentées plus tôt : des éléments de pure chance, des événements temporaires, des conditions fortuites, qui, à un moment donné, influencent les résultats. Les erreurs de mesure aléatoires peuvent être atténuées par divers moyens, soit en contrôlant directement à la source les aléas indésirables (principalement, par la standardisation et par des outils clairs et complets), soit en neutralisant les effets de ces aléas (par l'augmentation du nombre de mesures).

LES SOURCES D'ERREURS SYSTÉMATIQUES NE FONT PAS PARTIE DE LA FIDÉLITÉ

Les erreurs aléatoires ne sont pas les seules à influer sur les résultats obtenus à un instrument de mesure. Il y a aussi des **erreurs de mesure systématiques** ; ces erreurs sont constantes. Ce sont des biais qui ne varient pas d'une occasion à l'autre, qu'il s'agisse du même instrument que l'on fait passer à deux reprises, de deux formes équivalentes ou d'examinateurs différents. Voici divers exemples de facteurs pouvant causer ce type d'erreurs.

a) L'anxiété que ressentent toujours certains répondant lors d'un examen, peu importe la situation ou l'expérimentateur, a pour effet de nuire systématiquement à leurs résultats.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

b) Un test de personnalité ou un inventaire d'intérêts utilisé en contexte de sélection et dont l'objet mesuré est facilement détectable par les répondants, sera systématiquement affecté par la tendance de certains candidats à répondre de manière à se présenter sous un angle qu'ils croient désirables pour l'entreprise (p. ex., un candidat à un poste de vente aura tendance à se décrire comme sociable, persévérant, etc.).

c) Lors de leur formation, les examinateurs ont été induits en erreur sur la manière d'appliquer l'instrument de mesure et, à chaque fois, ils commettent la même erreur de chronométrage.

d) Les interviewers de cette organisation ont tous un préjugé fort tenace envers les employés provenant de tel secteur d'activité et aucun d'entre eux n'est parvenu à se faire embaucher.

e) Les rencontres d'évaluation se déroulent dans un contexte (p. ex., bureaux à aire ouverte) qui fait en sorte que les employés craignent systématiquement pour la confidentialité de leurs réponses. Il est alors plus difficile de les faire parler.

f) Lors d'une simulation, la procédure d'observation est intimidante pour les candidats et ils ont de la difficulté à se concentrer sur la tâche qui leur est demandée.

g) Le barème de correction à cet examen est beaucoup trop sévère, compte tenu des normes en vigueur dans la profession.

Les sources d'erreurs systématiques ne sont pas incluses dans la notion de fidélité. Elles sont stables et n'entraînent pas de différences de résultats d'une occasion à l'autre. Cependant, elles n'en constituent pas moins des sources importantes d'erreurs. Les moyens de les contrôler sont innombrables et doivent être adaptés à chaque situation. Ces moyens, pour la plupart, sont envisagés lors de la conception de l'instrument de mesure et de son application ; ils sont au cœur du concept de validité.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

DÉFINITIONS DE LA FIDÉLITÉ (R_{XX})

Au début de ce chapitre, la fidélité a été décrite par la **constance** ou la stabilité des résultats. Si l'on mesure la même caractéristique à plusieurs reprises, avec le même instrument ou avec des instruments comparables, et si les résultats obtenus sont toujours semblables, alors ils sont fidèles ; c'est la définition la plus courante. Il y en a une autre, plus précise sur le plan théorique et plus facile à comprendre ; cette fois, la fidélité est définie par la **précision** ou le degré d'exactitude des mesures. Si une caractéristique est mesurée de façon parfaitement exacte, sans aucune erreur aléatoire attribuable à l'anxiété du répondant, à la fatigue ou à tout autre facteur pouvant affecter sa performance, sans erreur fortuite attribuable à la fatigue de l'examinateur ou à ses fautes d'inattention, sans erreur attribuable à la situation changeante, à la formulation des questions, à l'imperfection des grilles de correction, alors les résultats auraient une fidélité parfaite. En d'autres termes, la fidélité est l'**absence d'erreurs aléatoires** (Guion, 1998 ; Kerlinger, 1986 ; Nunnally et Bernstein, 1994 ; *Standards*, 1999).

FIDÉLITÉ EN TANT QU'ABSENCE D'ERREURS ALÉATOIRES

Malgré tous les soins apportés par les spécialistes et autres responsables de l'application d'un instrument de mesure, les erreurs de mesure aléatoires ne sont jamais complètement éliminées ; aucun résultat n'atteint ce degré de perfection. En termes plus techniques, tout résultat obtenu (X) peut être vu comme étant constitué de deux composantes, un résultat vrai (v), sans erreur, et une erreur aléatoire (e), comme l'exprime la formule suivante (Spearman, 1910, cité dans Magnusson, 1966) :

Résultat obtenu (X) = $v + e$
où X : résultat obtenu
v : résultat vrai
e : erreur aléatoire

Le résultat vrai (v) est un idéal, une conception qui existe en théorie mais qui ne s'observe jamais dans la réalité. C'est le résultat qui serait obtenu si toutes les conditions, l'examinateur, le répondant

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

et l'instrument lui-même, étaient parfaites. Pour ce qui est de l'erreur aléatoire (*e*), elle traduit la somme des erreurs aléatoires de mesure. Avant de poursuivre, une légère modification s'impose à cette équation, qui est exacte seulement si toutes les erreurs sont aléatoires. Or, nous venons de voir que certaines erreurs ne sont pas aléatoires, mais plutôt systématiques et qu'elles affectent de façon constante tous les résultats à toutes les occasions (p. ex., préjugé généralisé chez les examinateurs ou barème de correction trop sévère). L'erreur systématique peut être considérée comme une constante qui s'ajoute « en plus » ou « en moins » au résultat vrai (p. ex., le résultat vrai de chaque candidat est amputé de cinq points à cause du barème de correction trop sévère). Pour tenir compte de cette réalité, il serait plus rigoureux de transformer l'équation de la façon suivante, où le résultat vrai (*v*) est remplacé par le résultat systématique (*s*), contenant à la fois le résultat vrai et l'erreur systématique (Guion, 1998) :

Résultat obtenu (X) = $s + e$
où X : résultat obtenu
s : résultat systématique, composé du résultat vrai et de l'erreur systématique
e : erreur aléatoire

Comparaison des concepts de fidélité et de validité. Cette équation, très riche sur le plan théorique, permet de rapprocher fidélité et validité. Un résultat **fidèle** est une mesure sans erreur aléatoire tandis qu'un résultat **valide** est une mesure sans erreur aléatoire ni erreur systématique : dans ce cas, le résultat obtenu correspondrait au résultat vrai, et seulement lui.

Lorsqu'un instrument de mesure est employé, c'est pour obtenir des résultats fiables, exacts et dignes de confiance. On est intéressé à ce que les erreurs de mesures soient réduites au minimum, pour obtenir un résultat le plus proche possible de la réalité (*v*). Car à quoi sert d'évaluer des candidats si les résultats qu'ils obtiennent ne traduisent pas exactement les compétences ou les caractéristiques recherchées. Bref, on désire des résultats à la fois valides et fidèles. La **validité** est obtenue principalement lors de l'élaboration de l'instrument de mesure et de son application ; de plus, il est possible d'évaluer

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

le niveau de validité pour un ensemble de données, par exemple, avec une démarche de validation basée sur la relation avec d'autres variables. Quant à la fidélité, elle dépend du contrôle des erreurs aléatoires réalisé grâce à divers moyens présentés au début de ce chapitre. Mais il reste à évaluer l'ampleur de la fidélité, à connaître jusqu'à quel point les résultats obtenus sont précis et à vérifier si les erreurs de mesure aléatoires ont effectivement été éliminées. D'un point de vue statistique, c'est une démarche simple à réaliser.

FIDÉLITÉ EN TANT QUE VARIANCE SYSTÉMATIQUE

D'abord, il faut savoir que l'erreur de mesure aléatoire (*e*) n'est pas corrélée au résultat systématique (*s*). Par exemple, prenons le cas d'un examinateur peu rigoureux qui commet de nombreuses fautes d'inattention. Que les réponses des candidats soient bonnes ou mauvaises, l'examinateur ne suit pas toujours la clé d'évaluation. Ses erreurs de correction varient de façon aléatoire, peu importe le niveau réel de performance du candidat ; elles sont indépendantes du score systématique. Cela permet de transformer l'équation précédente, qui s'applique à un seul résultat, en une équation valable pour un ensemble de résultats.

Variance totale (V_x) = V_s + V_e	
où V_x :	variance des résultats obtenus, ou variance totale
V_s :	variance des résultats systématiques, ou variance systématique
V_e :	variance des erreurs aléatoires, ou variance d'erreur

Cette nouvelle équation indique que, pour un ensemble de résultats, la variance totale des scores obtenus (V_x) est égale à la variance des résultats systématiques (V_s), plus la variance de l'erreur de mesure (V_e). Pour l'instant, seule la variance totale des scores obtenus (V_x) peut être calculée. Cependant, s'il n'y a aucune erreur aléatoire, la variance d'erreur (V_e) est nulle et la variance systématique (V_s) est égale à la variance totale (V_x). Mais il y a toujours de l'erreur aléatoire ; alors la variance systématique (V_s) est égale à la variance totale (V_x) moins la variance de cette erreur (V_e). Si l'on peut estimer la variance d'erreur (V_e), et cela est faisable, on pourra du coup connaître la variance systématique (V_s).

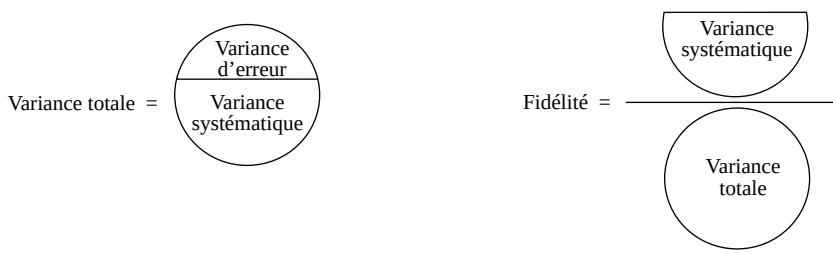
© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

C'est là l'essentiel du raisonnement à la base de l'estimation de la fidélité. Ayant été définie par l'absence d'erreur aléatoire, la fidélité correspond à la variance systématique (V_s), ou plutôt à la proportion de variance systématique (V_s) par rapport à la variance totale (V_x) :

Fidélité $\left(r_{xx}\right) = \dfrac{V_s}{V_x}$
où V_s : variance des résultats systématiques, ou variance systématique V_x : variance des résultats obtenus, ou variance totale

Ces explications peuvent être résumées par la figure suivante (voir figure 4.1).

Figure 4.1
VARIANCE ET FIDÉLITÉ



MÉTHODES D'ESTIMATION DE LA FIDÉLITÉ

Pour savoir si une mesure est exacte, on peut la comparer à d'autres mesures indépendantes du même objet. Si elles sont semblables, alors on conclut que la mesure est fiable ; en revanche, s'il y a des différences, on en déduit qu'il y a de l'erreur. De la même façon en gestion des ressources humaines, il est possible d'estimer la proportion d'erreur aléatoire à partir de deux ensembles de résultats mesurant la même chose. Il s'agit d'abord de mesurer deux fois le même objet chez les mêmes personnes, à l'aide du même instrument de mesure ou d'un instrument équivalent. Puis, il est facile de

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

calculer les proportions de variance systématique (V_s) et aléatoire (V_e) entre ces deux ensembles de mesures, généralement à l'aide du coefficient de corrélation.

Il existe plusieurs variantes de cette méthode pour estimer la fidélité, chacune d'elles reflétant différemment les diverses sources d'erreurs de mesure aléatoires. Les méthodes les plus fréquentes sont appelées test-retest, formes équivalentes, consistance interne et interexaminateurs.

ESTIMATION DE LA FIDÉLITÉ PAR TEST-RETEST

Mêmes personnes, même instrument de mesure, deux passations. La méthode test-retest consiste à faire passer aux mêmes personnes deux fois le même instrument de mesure. Par exemple, un test d'aptitude mentale générale est administré une première fois à un groupe de personnes. On attend, disons entre deux semaines et deux mois, le temps nécessaire pour que les répondants ne puissent se rappeler par cœur les réponses qu'ils ont données la première fois. Cependant, il ne faut pas trop attendre, pour éviter qu'il y ait possibilité de changement réel au niveau du trait mesuré (p. ex., une modification de l'aptitude). Après la période d'attente, le même test leur est administré une seconde fois. Chaque personne obtient ainsi deux résultats : un score obtenu lors de la première passation et un autre score au même test lors de la deuxième passation. Le tableau 4.2 illustre cette situation ; les résultats de 10 candidats fictifs suffisent à faire comprendre la méthode.

Pour estimer la fidélité, il s'agit d'évaluer dans quelle mesure les résultats de la deuxième passation sont reliés à ceux de la première. Les résultats n'ont pas à être identiques d'une passation à l'autre pour être parfaitement fidèles. En effet, il peut y avoir un écart explicable par l'effet de pratique ou de mémoire en faveur de la deuxième collecte. Par exemple, si toutes les personnes obtenaient exactement trois points de plus lors de leur deuxième passation, les mesures seraient tout de même parfaitement fidèles, car il n'y aurait aucune fluctuation aléatoire entre les deux passations. Les résultats obtenus à la deuxième passation seraient reliés à 100 % à ceux obtenus en première passation, parce qu'ils pourraient être prédits sans aucune erreur, simplement en ajoutant trois points au score original de chacun. De plus, le rang de chaque candidat par rapport aux autres demeurerait inchangé.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 4.2
ESTIMATION DE LA FIDÉLITÉ PAR LA MÉTHODE TEST-RETEST (EXEMPLE FICTIF)

	Première passation		Deuxième passation		
Sujet	**Résultat**	**Rang**	**Résultat**	**Rang**	**Différence de résultat**
1. Alain	119	1	120	1	+ 1
2. Brigitte	114	2	111	6	– 3
3. Carl	113	3	118	2	+ 5
4. Diane	112	4	112	5	0
5. Ernest	112	4	118	2	+ 6
6. Francine	111	6	107	7	– 4
7. Gilles	107	7	113	4	+ 6
8. Hélène	101	8	106	8	+ 5
9. Ian	101	9	96	10	– 5
10. Jeanne	96	10	100	9	+ 4
Moyenne	108,6		110,1		1,5
Écart type	6,8		7,5		4,1
Corrélation (*r*)	0,84				

Dans l'exemple du tableau 4.2, on peut voir que les mesures obtenues en deuxième passation sont reliées à celles de la première, sans toutefois être identiques. D'abord, en ce qui concerne le rang, les candidats n'ont pas subi de changement de plus de deux rangs, à l'exception de deux personnes (Brigitte et Gilles). Cependant, c'est en examinant la différence de résultats (dernière colonne) que l'on voit mieux l'ampleur des fluctuations : elles varient de moins cinq points (Ian) à plus six points (Ernest et Gilles). Il est intéressant de constater que, en moyenne, les personnes ont majoré leur score de 1,5 point (108,6 – 110,1), ce qui pourrait être attribuable à un effet de pratique : lors de la deuxième passation, les candidats étaient plus familiers avec ce genre de test et ont eu plus de facilité à s'y adapter. Bref, les deux ensembles de résultats sont reliés, mais pas de façon parfaite.

Calcul de la fidélité par le coefficient de corrélation. Il s'agit maintenant de quantifier précisément ce niveau de relation et ainsi estimer la fidélité des résultats. La corrélation produit-moment de

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Pearson (voir chapitre 7) est l'indice statistique le plus fréquemment utilisé à cet effet. Appliquée aux résultats de la première et de la deuxième passation, la corrélation obtenue est de 0,84 ; elle traduit, comme nous l'avions prévu, une relation forte entre les deux résultats (sachant que 1,0 est une corrélation parfaite alors que 0,0 est l'absence totale de relation). En outre, lorsqu'il s'agit de fidélité, la corrélation équivaut à la proportion de variance systématique par rapport à la variance totale, ce qui correspond à la définition de la fidélité. Il en est ainsi parce que c'est le même objet ou la même caractéristique qui est mesuré dans chaque ensemble de résultats. Dans ce cas, selon la théorie de la mesure, il est démontré que la corrélation n'a pas à être mise au carré pour obtenir le pourcentage de variance systématique. Donc, une corrélation de 0,84 indique que la variance systématique est estimée à 84 % et la variance d'erreur aléatoire, à 16 % ; l'estimation de la fidélité test-retest est de 0,84 pour les résultats de cet échantillon de 10 personnes à ce test d'aptitude.

Sources d'erreurs aléatoires prises en compte dans l'estimation. Les résultats ne sont pas complètement fidèles ; des fluctuations aléatoires sont observées d'une passation à l'autre, estimées à 16 % par le calcul de la corrélation. Dans le cas d'une méthode test-retest, ces fluctuations témoignent d'erreurs de mesure aléatoires à cause de changements incontrôlés (ou incontrôlables) survenus entre les deux collectes. Plusieurs sources d'erreurs, on l'a vu, peuvent influer sur le résultat d'un candidat. Certaines erreurs peuvent provenir du candidat lui-même, qui se trouve dans un état passager (p. ex., sa fatigue), a des réactions fortuites (p. ex., un trou de mémoire) ou est enclin à répondre au hasard (p. ex., lorsqu'il ne connaît pas une réponse). D'autres erreurs ont pour source l'examinateur, qui n'agit pas toujours de façon constante (p. ex., les fautes d'inattention), ou la situation, dont les conditions peuvent varier (p. ex., le bruit environnant). Enfin, l'instrument de mesure peut être à l'origine de ces fluctuations lorsque ses composantes souffrent d'un manque de standardisation ou de clarté.

Toutefois, l'estimation de la fidélité obtenue par test-retest ne parvient pas à détecter toutes les sources d'erreurs aléatoires. Par exemple, un problème d'échantillonnage des items ne sera pas mis en évidence parce que les items demeurent exactement les mêmes d'une passation à l'autre, présentant du coup les mêmes lacunes potentielles. Il en est de même pour la possibilité d'interaction entre

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

le candidat et l'examinateur : les deux personnes demeurent inchangées. Le fait que des sources d'erreurs ne soient pas prises en compte par la méthode test-retest a pour effet que l'estimation de la fidélité peut être plus élevée qu'elle n'aurait dû l'être en réalité. Le tableau 4.3 énumère les principales sources d'erreurs aléatoires incluses selon la méthode d'estimation employée.

L'estimation de la fidélité par la méthode test-retest donne lieu à un coefficient de **stabilité**, parce qu'elle vérifie jusqu'à quel point un groupe de personnes obtient les mêmes résultats à un même instrument auquel elles auraient été soumises à deux moments différents. La méthode test-retest est efficace seulement si les résultats obtenus en seconde passation ne peuvent pas être contaminés par la **mémoire**. Si les candidats peuvent se rappeler leurs réponses de la première passation et qu'ils utilisent ces souvenirs pour répondre en deuxième passation, alors la relation entre les deux résultats augmente artificiellement et le coefficient de fidélité est faussé à la hausse. Par exemple, la tendance à répondre au hasard comme source d'erreur aléatoire peut être neutralisée en partie par le fait qu'un candidat se souvient de certaines de ses réponses. On peut allonger la période de temps entre les deux passations pour atténuer l'effet possible de la mémorisation. Mais si la période est trop longue, un effet inverse peut se produire ; le coefficient de fidélité sera faussé à la baisse si des **changements** surviennent à la caractéristique mesurée. Par exemple, les répondants peuvent, entre les deux passations, acquérir de nouvelles connaissances ou changer. Il faut donc s'assurer que des changements importants pouvant affecter la caractéristique mesurée n'auront pas lieu au cours de la période couvrant la collecte des données.

ESTIMATION DE LA FIDÉLITÉ PAR FORMES ÉQUIVALENTES

Mêmes personnes, deux instruments de mesure, une ou deux passations. La méthode des formes équivalentes permet de contourner cette double difficulté de la mémorisation et du changement de la caractéristique mesurée. La démarche pour estimer la fidélité est semblable à celle du test-retest, sauf que c'est une forme équivalente de l'instrument de mesure qui est administrée lors de la deuxième passation. Une version équivalente, appelée aussi forme parallèle, est une deuxième version de l'instrument de mesure qui, sans être identique, possède les mêmes

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 4.3

SOURCES D'ERREURS DE MESURE PRISES EN COMPTE PAR LES DIVERSES MÉTHODES D'ESTIMATION DE LA FIDÉLITÉ

Sources d'erreurs	1 Test-retest avec délai	2 Équivalence sans délai	3 Équivalence avec délai	4 Consistance interne	5 Interexaminateurs
A) Le candidat					
1. Tendance à répondre au hasard.	Exclu si mémorisation	Inclus	Inclus	Inclus	
2. État général mais passager du candidat.	Inclus	Inclus si changement entre les passations	Inclus	Inclus si changement durant la passation	Inclus si changement durant la passation
3. Réactions spécifiques fortuites ou inhabituelles du candidat.					
a) compréhension des directives,	Exclu si mémorisation	Inclus	Inclus		
b) trous de mémoire,	Inclus	Inclus	Inclus	Inclus	
c) fautes d'inattention.	Inclus	Inclus	Inclus	Inclus	
4. Interaction avec les caractéristiques de l'examinateur.					Inclus

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

B) L'examinateur					
5. État général mais passager de l'examinateur.	Inclus	Inclus si changement entre les interventions	Inclus	Inclus si changement durant l'intervention	Inclus
6. Réactions spécifiques fortuites ou inhabituelles de l'examinateur.	Inclus	Inclus	Inclus	Inclus	Inclus
7. Interaction avec les caractéristiques du candidat.					Inclus
C) La situation					
8. Conditions d'administration non standardisées.	Inclus	Inclus si changement entre les passations	Inclus	Inclus si changement durant la passation	Inclus si changement durant la passation
D) L'instrument					
9. Composantes non standardisées					
a) items,	Inclus	Inclus	Inclus	Sans objet	
b) directives,	Inclus	Inclus	Inclus		
c) outils d'évaluation.	Inclus	Inclus	Inclus	Inclus si changement pour un même répondant	Inclus

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 4.3 *(suite)*

SOURCES D'ERREURS DE MESURE PRISES EN COMPTE PAR LES DIVERSES MÉTHODES D'ESTIMATION DE LA FIDÉLITÉ

Sources d'erreurs	1 Test-retest avec délai	2 Équivalence sans délai	3 Équivalence avec délai	4 Consistance interne	5 Interexaminateurs
10. Ambiguïtés et insuffisances					
a) items,	Exclu si mémorisation	Inclus	Inclus	Inclus	
b) directives,	Exclu si mémorisation	Inclus	Inclus		
c) outils d'évaluation.	Exclu si mémorisation	Inclus	Inclus	Inclus si changement d'interprétation pour un même répondant	Inclus
11. Échantillonnage des items (*content sampling*)		Inclus	Inclus	Inclus	

Note : Pour les méthodes d'estimation nº 1 à nº 4, on présume qu'il y a un seul examinateur, alors que pour la méthode nº 5, il s'agit du même instrument de mesure administré une seule fois.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

propriétés métriques : elle mesure les mêmes contenus, avec le même type d'items, le même degré de difficulté, de sorte que la distribution des scores est semblable à la forme originale de l'instrument.

Par exemple, une version A et une version B d'un test d'aptitude sont administrées à un même groupe de personnes. Selon une première variante, les deux versions peuvent être administrées successivement en **une seule passation**, sans délai entre les deux. Procéder ainsi permet de mieux contrôler la standardisation des conditions d'administration et de certains états passagers des répondants et de l'examinateur (humeur, santé ou préparation mentale). Par contre, si la nature du test fait craindre des effets probables de fatigue ou de pratique, il vaut mieux prévoir un délai entre les deux versions, de deux semaines à deux mois par exemple. Il y a alors **deux passations**. Pour cette seconde variante, on doit envisager la possibilité qu'il y ait des changements chez le répondant en ce qui a trait à la caractéristique mesurée.

Calcul de la fidélité par le coefficient de corrélation. Le tableau 4.4 présente des données (fictives) recueillies en suivant la méthode des formes parallèles. Chaque personne obtient d'abord un premier résultat à la version A de l'instrument de mesure. Pour exemplifier, supposons qu'il s'agit du même test d'aptitude mentale que celui utilisé par la méthode test-retest, présenté à la section précédente, et que les résultats recueillis pour la forme A sont ceux provenant de la première passation (voir tableau 4.2). Chaque personne a également un autre résultat obtenu cette fois à la version B du test, qui est une forme équivalente à la version A. Le reste de la démarche est analogue à celle de la méthode test-retest. Si l'on compare les données d'une forme à l'autre, on constate que, malgré des fluctuations observables, il existe un certain niveau de relation. Par exemple, aucun candidat n'enregistre un changement supérieur à trois rangs. En ce qui concerne la différence de résultats, la variation est contenue entre plus six et moins sept points. Ce niveau de relation se reflète dans la corrélation de 0,79, qui traduit 79 % de variance systématique et 21 % de variance d'erreur aléatoire.

Sources d'erreurs aléatoires prises en compte dans l'estimation. Comparée à l'estimé de 0,84 produit par la méthode test-retest, l'estimation de 0,79 obtenue par les formes équivalentes est plus faible. Comment expliquer cette différence ? Bien sûr, les données sont fictives, mais elles servent à illustrer deux principes. Premièrement,

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 4.4
ESTIMATION DE LA FIDÉLITÉ PAR LA MÉTHODE DES FORMES ÉQUIVALENTES (EXEMPLE FICTIF)

Sujet	Forme A		Forme B		
	Résultat	Rang	Résultat	Rang	Différence de résultat
1. Alain	119	1	118	1	– 1
2. Brigitte	114	2	117	2	+ 3
3. Carl	113	3	108	5	– 5
4. Diane	112	4	108	5	– 4
5. Ernest	112	4	117	2	+ 5
6. Francine	111	6	104	8	– 7
7. Gilles	107	7	113	4	+ 6
8. Hélène	101	8	102	9	+ 1
9. Ian	101	9	107	7	+ 6
10. Jeanne	96	10	96	10	0
Moyenne	108,6		109,0		0,4
Écart type	6,8		6,9		4,4
Corrélation (*r*)	0,79				

des méthodes différentes ont tendance à engendrer des estimations de valeurs différentes, pour la simple raison que ces méthodes ne vérifient pas nécessairement la présence des mêmes sources d'erreurs[2]. Deuxièmement, *plus nombreuses sont les sources d'erreurs dont une méthode tient compte, plus les estimations de la fidélité qu'elle génère auront tendance à diminuer.* Rappelons que la fidélité est définie par rapport à la variance systématique, qui est égale à la variance totale moins la variance d'erreur aléatoire. Donc, plus il y a de sources d'erreurs qui peuvent affecter la variance aléatoire, plus cette variance sera grande et moins, réciproquement, il « restera » de variance systématique. Or, la méthode des formes équivalentes considère généralement plus de sources d'erreurs que la méthode test-retest. Par exemple, l'erreur due à

2. Une autre raison est l'erreur d'échantillonnage dont la probabilité est très importante avec un échantillon aussi petit que 10 personnes. Il ne s'agit là que d'un exemple : une étude de fidélité exige un échantillon beaucoup plus grand et plus représentatif de la population.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

l'échantillonnage des items influence les résultats lorsque les deux tests sont des versions équivalentes. Les items n'étant plus exactement les mêmes d'une version à l'autre, ils peuvent être la source de variations indésirables dans les résultats. De plus, si l'on a recours à deux passations séparées par un délai, la méthode des formes équivalentes inclut pratiquement toutes les sources d'erreurs, du moins toutes celles prises en compte par la méthode test-retest. Le tableau 4.3 permet de poursuivre ainsi la comparaison entre les deux méthodes.

La fidélité obtenue par la méthode des formes équivalentes est appelée coefficient d'**équivalence**. Cette méthode est appropriée si les formes sont vraiment équivalentes et si les résultats à la seconde forme ne peuvent pas être influencés par la mémoire ou par l'apprentissage des répondants. De plus, lorsqu'il y a un délai entre les collectes, il ne doit pas y avoir de changements pouvant influencer la caractéristique mesurée.

ESTIMATION DE LA FIDÉLITÉ PAR CONSISTANCE INTERNE

Mêmes personnes, même instrument de mesure, une passation. L'estimation par consistance interne est une méthode largement utilisée. Des plus pratiques, elle ne requiert qu'une seule passation et une seule forme de l'instrument de mesure. Fondamentalement, la consistance interne indique jusqu'à quel point les diverses parties d'un instrument (items, questions ou problèmes) mesurent la même chose (Guion, 1965). Imaginons un examen de mathématiques ne comprenant que des additions simples (p. ex., 34 + 17). Normalement, une personne qui sait répondre à l'un de ces problèmes devrait, en principe pouvoir résoudre les autres. Si tel est le cas, on dit de cet instrument de mesure qu'il a de la consistance interne ou qu'il est homogène (Gatewood et Feild, 1998 ; Guion, 1998). Lorsqu'un instrument mesure une seule dimension, comme cet examen composé seulement de problèmes d'addition, il est qualifié d'**unidimensionnel** ; lorsqu'il mesure plus d'une dimension (p. ex., addition, soustraction, multiplication et division), il est **multidimensionnel**.

Pour estimer le niveau de fidélité en se basant sur la consistance interne, il faut examiner les résultats entre des parties équivalentes de l'instrument. Plus la corrélation entre les parties est forte, c'est-à-dire plus les réponses des candidats à une partie de l'instrument sont reliées à celles des autres parties, moins il y a d'erreurs aléatoires et

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

plus les résultats sont fidèles. Il y a différentes méthodes pour calculer le coefficient de consistance interne ; les plus fréquentes sont la bissection et le coefficient alpha (α).

Bissection. La bissection consiste d'abord à appliquer l'instrument de mesure à un échantillon de répondants. Après, l'instrument est divisé en deux moitiés équivalentes, de manière à pouvoir compiler un résultat pour chaque moitié et à établir ensuite la corrélation entre ces deux ensembles de résultats. Reprenons l'exemple du test d'aptitude et supposons qu'il compte 50 questions. Une fois administré, il est possible de le scinder en deux parties de 25 questions chacune, puis de procéder comme pour les formes équivalentes. Pour que cette méthode soit valable, cependant, il faut que les deux moitiés de l'instrument soient équivalentes : chacune d'elles doit mesurer le même contenu et donner lieu à des résultats dont les distributions sont comparables. Sinon, il y aura des fluctuations dans les résultats d'une moitié à l'autre qui viendront gonfler la proportion d'erreurs aléatoires. Dans l'exemple du test d'aptitude mentale, les énoncés mesurent le même contenu (test unidimentionnel), mais ils apparaissent dans le test par ordre croissant de difficulté. Il est alors habituel de constituer les deux moitiés équivalentes en répartissant de part et d'autre les 25 énoncés pairs et les 25 impairs ; de cette façon, on est assuré d'avoir deux moitiés de test de difficulté égale.

Il reste à compiler le résultat à chaque moitié pour chacun des répondants, ce qui est fait au tableau 4.5. Encore une fois, les deux ensembles de résultats (énoncés pairs et impairs) sont passablement reliés. Les rangs ont bougé d'au plus trois unités alors que la différence de résultats n'excède pas trois points en plus ou en moins. L'ampleur de la relation est estimée à 0,85 par le calcul du coefficient de corrélation. Cependant, ce coefficient ne représente pas la fidélité de l'instrument complet avec ses 50 énoncés, mais plutôt la fidélité de chacune des moitiés ne comportant que 25 énoncés. Or, suivant le principe que les erreurs aléatoires ont tendance à s'annuler par l'augmentation du nombre d'énoncés, un instrument de 50 items sera plus fidèle qu'un instrument de 25 items, toutes choses égales d'ailleurs. Pour obtenir la fidélité de l'instrument complet, il faut corriger la corrélation obtenue en appliquant la formule Spearman-Brown. L'application de cette formule permet d'estimer à 0,92 la fidélité du test d'aptitude complet (voir tableau 4.5).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Spearman-Brown $\left(r_{xx}\right) = \dfrac{2\,r}{1+r}$
où r_{xx} : fidélité pour la pleine longueur de l'instrument de mesure
r : corrélation entre les deux moitiés de l'instrument de mesure

Tableau 4.5
ESTIMATION DE LA FIDÉLITÉ PAR LA MÉTHODE DE LA BISSECTION (EXEMPLE FICTIF)

Sujet	**25 items pairs**		**25 items impairs**			**50 items**
	Résultat	**Rang**	**Résultat**	**Rang**	**Différence de résultat**	**Résultat**
1. Alain	59	1	60	1	+ 1	119
2. Brigitte	56	4	58	2	+ 2	114
3. Carl	58	2	55	5	– 3	113
4. Diane	55	5	57	3	+ 2	112
5. Ernest	57	3	55	5	– 2	112
6. Francine	54	6	57	3	+ 3	111
7. Gilles	54	6	53	7	– 1	107
8. Hélène	52	8	49	10	– 3	101
9. Ian	51	9	50	8	–1	101
10. Jeanne	48	10	48	9	0	96
Moyenne	54,4		54,2		– 0,2	
Écart type	3,2		3,9		2,0	
Corrélation (r)	0,85					

Note : Spearman-Brown : $r_{xx} = (2\,r) \div (1 + r) = (2 \times 0{,}85) \div (1 + 0{,}85) = 0{,}92$

Sources d'erreurs aléatoires prises en compte dans l'estimation. L'estimation de la fidélité par consistance interne, dont la bissection est une variante, repose sur une seule passation. Par conséquent, elle ne peut vérifier si les résultats sont empreints d'erreurs de mesure causées par des états passagers du candidat ou de l'examinateur, ou des changements survenus dans le temps dans les conditions d'administration ou les outils d'évaluation. À moins de fluctuations durant l'intervention, il est impossible d'évaluer l'effet de ces facteurs sur les résultats

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

(voir tableau 4.3). Un changement de directives d'un candidat à l'autre passera également inaperçu, car, pour chaque candidat, les directives demeurent constantes au cours de leur unique passation. Quant aux items, la question de leur changement ne se pose pas vraiment ; comme les approches par consistance interne sont généralement réservées aux tests psychométriques ou aux examens standardisés, les items sont toujours constants pour ces instruments de mesure.

Cependant, la fidélité obtenue par **consistance interne**, appelée aussi coefficient de consistance interne, permet d'évaluer l'influence des erreurs aléatoires ayant pour sources la formulation des items (ambiguïtés et insuffisances) et leur échantillonnage. Elle vérifie aussi la présence d'erreurs momentanées de la part du répondant ou de l'examinateur, comme les fautes d'inattention et les trous de mémoire. S'il existe une possibilité que les candidats répondent au hasard, elle sera également incluse dans l'estimation de la fidélité.

Coefficient alpha (α). La fidélité estimée par la bissection est basée sur la division de l'instrument de mesure en deux moitiés équivalentes. Si tous les items mesurent la même chose avec les mêmes qualités métriques, l'estimation de la fidélité devrait être la même, peu importe la façon dont les items seront répartis. Signalons que les items ne sont jamais tous parfaitement équivalents, que ce soit au regard du contenu, de la formulation, du degré de difficulté, etc. De plus, il y a l'erreur causée par l'échantillonnage des items, toujours possible. L'estimation de la fidélité pourra donc changer selon la répartition des items dans les deux moitiés. Pour contourner cette difficulté, on peut calculer la fidélité de toutes les bissections imaginables, puis en compiler la moyenne. Mais il n'est pas nécessaire de suivre un aussi long procédé pour obtenir cette moyenne : il existe une formule relativement simple, appelée alpha (α) de Cronbach, qui évalue la proportion de l'ensemble des intercorrélations entre tous les items par rapport à la variance totale (voir Cronbach, 1990)[3]. Comme l'application de cette formule peut rapidement devenir fastidieuse lorsque l'instrument compte plusieurs items, on s'en remet à des logiciels statistiques.

3. Dans le cas d'items dichotomiques dont la réponse est bonne ou mauvaise, on peut utiliser la formule Kuder-Richardson-20 (voir Cronbach, 1990).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Étant basée sur la degré de consistance interne entre les items, cette approche ne fournit pas une bonne estimation de la fidélité dans le cas d'un instrument **multidimensionnel** (p. ex., addition, soustraction, multiplication et division). Prenons l'exemple d'un examen de connaissances, dont les items auraient été choisis pour mesurer diverses dimensions du domaine informatique. L'indice de consistance interne pour un tel examen serait assez faible, ce qui traduirait le degré d'hétérogénéité des items, et pas nécessairement la présence d'erreurs aléatoires. Il y a au moins deux solutions à ce problème. Premièrement, s'il y a suffisamment d'items par dimension, on peut diviser l'instrument en autant de parties (unidimensionnelles), puis calculer le coefficient alpha pour chacune d'elles ; des formules adaptées à cette situation sont également disponibles (voir Cronbach, 1990). Deuxièmement, on peut toujours recourir à la bissection, à condition de pouvoir répartir les items de manière à obtenir deux parties équivalentes.

La consistance interne est trompeuse pour les tests de vitesse. Que ce soit par bissection ou par coefficient alpha (α), la consistance interne n'est pas non plus une méthode appropriée dans le cas des instruments comportant beaucoup de vitesse. Dans le domaine de la mesure, on distingue deux types d'instruments : les tests de vitesse et les tests de puissance. Il est généralement reconnu qu'un test de **puissance** est un test que la plupart des personnes ont le temps de compléter ou au moins ont le temps d'essayer tous les items qu'il renferme ; un examen scolaire constitue un exemple de test de puissance. D'un autre côté, un test de **vitesse** comporte une limite de temps donné et la plupart des personnes n'ont pas le temps de le terminer ; un test d'aptitude est un exemple typique.

Quand un instrument comporte ainsi un facteur vitesse, il est facile de comprendre pourquoi la consistance interne n'est pas une bonne façon d'estimer la fidélité. D'abord, dans un test de vitesse, c'est principalement le nombre d'items auxquels elles ont répondu qui différencie les personnes. Ces dernières répondent souvent correctement à la plupart des items auxquels elles se rendent, de sorte que les résultats obtenus à ces items sont fortement corrélés entre eux. Ensuite, les résultats sont aussi fortement corrélés entre les items restants. N'ayant pas eu le temps d'y répondre, plusieurs personnes échouent forcément ces items. Par conséquent, si l'on utilise une bissection de type pair-impair, le coefficient de fidélité sera grandement

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

gonflé parce que les scores aux items pairs et aux items impairs seront pratiquement identiques pour chaque répondant, et le coefficient de fidélité frôlera la perfection ; le même phénomène inflationniste se répétera pour l'indice alpha. Par ailleurs, si l'on désire calculer une bissection entre les scores obtenus à la première moitié du test et ceux obtenus à la seconde moitié, la plupart des répondants auront pratiquement tous les items corrects à la première moitié, alors qu'il y aura des variations entre eux à la seconde moitié ; il en résultera un coefficient de fidélité relativement faible.

La meilleure façon d'estimer la fidélité d'un test de vitesse est de construire deux formules équivalentes. Si c'est impossible, on peut recourir à une bissection de type pair-impair, mais seulement si les deux moitiés sont administrées et chronométrées séparément. La méthode par test-retest peut être utilisée, mais au risque d'être contaminée par les effets de la mémoire et de la pratique.

ESTIMATION DE LA FIDÉLITÉ INTEREXAMINATEURS

Mêmes personnes, même instrument de mesure, une passation, deux examinateurs ou plus. De nombreux instruments de mesure employés en gestion des ressources humaines nécessitent la participation d'examinateurs pour l'observation ou l'évaluation des personnes ; il en est ainsi pour l'appréciation du rendement de ses subordonnés, la classification d'un emploi, la correction d'un examen comportant des questions ouvertes ou l'évaluation d'un candidat à l'aide d'une entrevue ou d'une simulation. On dit de ces instruments qu'ils sont **subjectifs**, en raison du rôle important que le jugement des examinateurs joue dans les résultats obtenus. Il y a alors de fortes chances que le résultat obtenu par l'évalué ne dépende pas seulement de sa performance, mais aussi de l'évaluateur : de son humeur, de son manque d'attention, de ses préjugés, de sa façon de comprendre comment appliquer les outils d'évaluation et aussi de la consistance avec laquelle il applique ces outils. L'influence de l'examinateur peut être évaluée en comparant deux ou plusieurs examinateurs lorsqu'ils évaluent la même personne. S'il y a concordance entre les examinateurs, les résultats sont fidèles ; sinon, il y a erreur de mesure. Pour éviter que d'autres sources d'erreurs ne se manifestent, on a généralement recours au même instrument de mesure administré une seule fois. Par exemple, un examen est corrigé de façon

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

indépendante par deux évaluateurs différents ou un candidat passe une entrevue avec plusieurs interviewers, qui évaluent chacun pour soi la performance du candidat.

Calcul de la fidélité par le coefficient de corrélation. Le cas le plus simple repose sur deux examinateurs. Prenons l'exemple d'un test situationnel, l'épreuve du courrier, administré à 10 candidats dans le cadre d'un concours de promotion. Pour chacun des 25 items que comporte le test, le candidat devait indiquer, par écrit, ce qu'il ferait s'il se trouvait dans cette situation. Les réponses, très variées, étaient alors corrigées à l'aide d'un solutionnaire détaillé comprenant les éléments de réponse attendus et leur valeur respective ; le test était sur 60 points. Le tableau 4.6 présente les résultats réels obtenus lorsque deux examinateurs différents ont corrigé à l'aveugle, sans se consulter, les réponses des candidats. Étant donné le caractère exhaustif du solutionnaire et la grande standardisation de la démarche, on observe une convergence étroite entre les examinateurs : à une exception près, la différence ne dépasse pas trois points. La fidélité interexaminateurs est estimée à 0,96 par le calcul de la corrélation. Notons que cette méthode peut faire intervenir plus de deux examinateurs (Gatewood et Feild, 1994).

Sources d'erreurs aléatoires prises en compte dans l'estimation. La fidélité interexaminateurs, appelée aussi interjuges, évalue l'influence de toutes les sources d'erreurs aléatoires reliées à l'examinateur, que ce soit son état passager, ses réactions fortuites ou les interactions avec le répondant (voir tableau 4.3). Un manque de standardisation ou des ambiguïtés dans les outils d'évaluation font aussi partie des sources d'erreurs considérées. Basée sur une seule passation, cette méthode ne peut détecter l'effet des changements dans le temps, à moins que ces derniers ne prennent place durant la passation, pas plus que la tendance du candidat à répondre au hasard et ses réactions fortuites.

La fidélité intra-examinateur. Une autre façon de vérifier si l'examinateur est une source d'erreur de mesure consiste à examiner la constance chez un même examinateur d'une personne évaluée à l'autre. Autrement dit, il y a fidélité intra-examinateur lorsque la même performance ou le même élément de réponse amène toujours la même évaluation de la part du même évaluateur. Par exemple, deux réponses semblables à un examen fournies par deux étudiants

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 4.6
ESTIMATION DE LA FIDÉLITÉ INTEREXAMINATEURS

Sujet	Examinateur A		Examinateur B		
	Résultat	Rang	Résultat	Rang	Différence de résultat
1.	24	4	23	3	– 1
2.	19	8	17	8	– 2
3.	12	10	12	10	0
4.	24	4	21	5	– 3
5.	23	6	16	9	– 7
6.	22	6	20	6	– 2
7.	25	3	23	3	– 2
8.	17	9	19	7	+ 2
9.	40	1	39	1	– 1
10.	33	2	30	2	– 3
Moyenne	23,9		22,0		– 1,9
Écart type	7,9		7,7		2,3
Corrélation (*r*)	0,96				

devraient recevoir la même note de leur professeur. Deux candidats ayant les mêmes qualifications devraient être évalués au même niveau par l'interviewer, peu importe la séquence des rencontres ou le moment de la journée. Les ouvrages spécialisés ne proposent pas de méthode pour estimer cet aspect néanmoins important de la fidélité. Plusieurs avenues sont envisageables. Par exemple, on demande à un examinateur de corriger deux fois le même instrument, en prenant toutefois des moyens pour éliminer la possibilité de mémorisation. Ou encore, il est possible de vérifier si le même élément de réponse, qui se répète chez un même candidat ou d'un candidat à l'autre, reçoit toujours le même score.

COMPARAISON DES DIVERSES MÉTHODES D'ESTIMATION

Toutes les méthodes d'estimation présentées poursuivent le même but : fournir une estimation de la fidélité, c'est-à-dire la proportion de variance qui n'est pas de l'erreur de mesure aléatoire. Toutefois, si l'on

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

se reporte au tableau 4.3, on remarque que chacune de ces méthodes est différente pour ce qui est des sources d'erreurs aléatoires considérées dans leur calcul. Par conséquent, bien que ces diverses méthodes visent à évaluer la présence d'erreurs aléatoires en comparant des mesures du même objet, les coefficients de fidélité produits ne sont pas interchangeables (*Standards*, 1999). La méthode test-retest vérifie l'effet possible des changements qui pourraient survenir dans le temps ; le coefficient fidélité obtenu traduit principalement la **stabilité** des résultats pour une période de temps donnée. La méthode des formes équivalentes, sensible aux différences entre deux formes du même instrument de mesure, produit un coefficient d'**équivalence**. La méthode par consistance interne évalue les différences entre des items au contenu analogue et appartenant au même instrument de mesure ; le coefficient de fidélité estime la **consistance** entre ces items. Finalement, analyser les divergences entre examinateurs permet de calculer la fidélité **interexaminateurs**.

Guion (1998) recommande de distinguer les méthodes et de retenir celle qui est la plus appropriée aux circonstances, le meilleur indice de fidélité étant celui qui tient compte des erreurs les plus préoccupantes. Il faut se poser la question suivante : est-ce que cet indice particulier de fidélité néglige de considérer des sources d'erreurs potentiellement importantes (Cronbach, 1990) ? Par exemple, si un instrument est utilisé pour faire des prédictions à long terme, la stabilité de la mesure devient très importante. Ou encore, si l'on a recours à des examinateurs différents pour corriger un examen, l'équité du processus ne pourra être assurée sans la fidélité entre ces examinateurs. En contrepartie, il ne faudra pas oublier que, dans ces deux cas, ni la fidélité test-retest, ni la fidélité interexaminateurs ne fournit d'indication sur la possibilité d'erreur due à l'échantillonnage des items.

En résumé, il faut penser à plusieurs coefficients de fidélité distincts, selon les sources d'erreurs considérées[4].

4. C'est là une prémisse de la théorie de la généralisation, qui constitue une approche plus globale et plus intégrée pour estimer la fidélité. À l'instar du devis expérimental traditionnel et de son pendant statistique qu'est l'analyse de variance, la théorie de la généralisation distingue les diverses sources d'erreurs et tente simultanément d'évaluer l'influence de chacune d'elles sur les résultats (voir Cronbach, 1990 ; Nunnally et Bernstein, 1994 ; *Standards*, 1999).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

USAGES PRATIQUES DU COEFFICIENT DE FIDÉLITÉ

Le premier usage du coefficient de fidélité est de fournir une estimation du degré de précision, de fiabilité des résultats obtenus à un instrument de mesure. En soi, la fidélité est une indication de la qualité des résultats en ce qui concerne la présence ou l'absence d'erreurs de mesure aléatoires. Le coefficient de fidélité peut aussi servir à d'autres fins, et en voici deux particulièrement importantes.

PLAFOND SUR LA VALIDITÉ ET CORRECTION POUR ATTÉNUATION

La fidélité est un plafond à la validité. Nous savons que la validité est la qualité la plus fondamentale d'une mesure ; elle indique jusqu'à quel point les résultats mesurent ce qu'ils sont censés mesurer ou jusqu'à quel point les inférences faites à partir des résultats sont exactes. Du coup, la fidélité devient aussi une qualité importante, parce qu'elle impose un plafond à la validité (Guion, 1998 ; *Standards*, 1999). En effet, la présence d'erreurs aléatoires atténue la validité, de sorte que cette dernière ne peut excéder la racine carrée de la fidélité (Cronbach, 1990). Par exemple, si la fidélité est de 0,80 pour un ensemble de résultats, alors 0,89 (soit la racine carrée de 0,80) est la limite supérieure que peut atteindre la validité de ces résultats. Toutefois, une fidélité élevée ne garantit pas la validité des résultats : la fidélité est une condition nécessaire, mais non suffisante. Kerlinger (1973) disait de la fidélité, qu'elle était comme l'argent : c'est d'en manquer qui pose problème.

Formules de correction pour atténuation. La relation théorique entre la fidélité et la validité est exprimée par la formule suivante (Guion, 1998 ; Nunnally et Bernstein, 1994) :

Validité si critère et prédicteur parfaitement fidèles $\left(r_{x\infty\, y\infty}\right) = \dfrac{r_{xy}}{\sqrt{r_{xx}\, r_{yy}}}$

où r_{xy} : validité observée entre le prédicteur et le critère

r_{xx} : fidélité observée du prédicteur (x)

r_{yy} : fidélité observée du critère (y)

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Cette formule permet de calculer la corrélation entre deux variables, par exemple un prédicteur et un critère, si la fidélité de ces variables était parfaite. Le fait d'estimer cette corrélation idéale est appelé correction pour atténuation due à l'infidélité. Supposons que la corrélation observée entre deux variables est de 0,40 (validité critériée) et que la fidélité de chacune d'elles est respectivement de 0,80 et de 0,60. Alors, si on avait pu mesurer ces deux variables de façon parfaitement fidèle, la validité aurait été de 0,58 (soit $0{,}40 \div \sqrt{0{,}80 \times 0{,}60}$). Si l'on désire faire la correction pour une seule variable, il suffit de ne retenir que cette variable dans la formule. Par exemple, si la correction n'est effectuée que pour le prédicteur, alors la validité devient 0,45 (soit 0,40 divisé par $\sqrt{0{,}80}$).

D'une certaine manière, l'usage de cette formule peut être trompeur. À quoi bon rêver à un niveau de validité hypothétique si, dans les faits, il est impossible d'obtenir des mesures parfaitement fidèles. Il serait donc plus approprié de ne recourir à la correction pour atténuation que si l'on peut réellement envisager une certaine amélioration de la fidélité du prédicteur ou du critère. Cela peut être calculé par la formule suivante (Nunnally et Bernstein, 1994) :

Validité si amélioration de la fidélité du critère et du prédicteur

$$\left(r'_{xy}\right) = r_{xy} \frac{\sqrt{r'_{xx}\, r'_{yy}}}{\sqrt{r_{xx}\, r_{yy}}}$$

où r_{xy} : validité observée entre le prédicteur et le critère

r'_{xx} : fidélité améliorée du prédicteur (x)

r'_{yy} : fidélité améliorée du critère (y)

r_{xx} : fidélité observée du prédicteur (x)

r_{yy} : fidélité observée du critère (y)

Reprenons les données de l'exemple précédent et supposons qu'il soit réaliste d'améliorer la fidélité du prédicteur de 0,80 à 0,90 et celle du critère de 0,60 à 0,70. La validité obtenue de 0,40 passerait alors à une validité corrigée de 0,46 (soit $0{,}40 \times [(\sqrt{0{,}90 \times 0{,}70}) \div (\sqrt{0{,}80 \times 0{,}60})]$). Advenant que l'on puisse seulement améliorer la fidélité du prédicteur, alors la validité corrigée serait de 0,42 (soit $0{,}40 \times [\sqrt{0{,}90} \div \sqrt{0{,}80}]$). Cette version de la formule est

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

utile pour estimer la fluctuation de la validité si des améliorations sont apportées à la fidélité des mesures. Lorsqu'on regarde les exemples, on se rend compte que les changements ne sont pas toujours si grands.

Ces formules sont bien belles en théorie, mais leur application n'est pas si simple en pratique. Il faut se rappeler que la fidélité exacte n'est jamais connue, mais toujours estimée, et que chaque estimation n'est pas interchangeable en raison des erreurs aléatoires considérées. Alors, quelle estimation faut-il utiliser dans ces formules ? Exemple à l'appui, Dunnette (1966) rappelle depuis longtemps les vertus de la prudence et l'importance de bien connaître la nature des variables en cause avant de tirer des conclusions.

ERREUR TYPE DE LA MESURE (S_e) ET INTERVALLE DE CONFIANCE

Calcul de l'erreur type de la mesure. La fidélité est un indice de précision qui s'applique à un ensemble de résultats et il n'indique pas le degré de précision, ou la marge d'erreur, à accorder à un résultat en particulier. Par exemple, une personne se plaint du score de 105 qu'elle a obtenu à un test d'aptitude mentale et allègue que, n'eût été de sa fatigue momentanée (source d'erreur aléatoire), elle aurait pu avoir de 10 à 15 points de plus. Est-il possible que cette personne ait raison, sachant que la fidélité des résultats à ce test est estimée 0,90 par la méthode test-retest ? Le concept d'erreur type de la mesure permet de répondre à cette question et d'établir un intervalle de confiance autour d'un résultat donné. La formule est la suivante :

Erreur type $(S_e) = S_x \sqrt{1 - r_{xx}}$
où S_x : écart type des résultats observés
r_{xx} : fidélité des résultats observés

À supposer que l'écart type des résultats de la population à ce test d'aptitude soit de 20, on obtient une erreur type égale à 6,3 (soit $20 \sqrt{1 - 0{,}90}$). L'erreur type est une estimation de l'écart type de la distribution des erreurs aléatoires autour du score vrai d'une personne. Autrement dit, imaginons que cet instrument de mesure, ou plusieurs formes du même instrument, soit administré au même individu à d'innombrables reprises. Comme les résultats ne sont pas parfaitement fidèles, cet individu obtiendrait de nombreux résultats

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

différents, plus ou moins proches les uns des autres. L'erreur type est une estimation de l'écart type de cette distribution ; il traduit l'étendue de l'erreur de mesure pour cette personne. Quant à la moyenne des résultats obtenus, elle correspond à son score vrai, au résultat qu'elle aurait s'il n'y avait pas d'erreur aléatoire[5].

Il est recommandé de toujours calculer l'erreur type des instruments que l'on utilise. Il vaut mieux également calculer l'erreur type pour chacun des sous-groupes ayant répondu à l'instrument. En effet, si un instrument présente la même fidélité pour deux groupes de sujets, disons 0,75, mais que l'écart type des scores du premier groupe est de 14, tandis que l'écart type du second est de 7, les erreurs types correspondantes seront substantiellement différentes, comme on peut le voir à l'aide des calculs ci-dessous. On accordera donc plus de confiance à un score obtenu par un sujet du deuxième groupe qu'à celui obtenu par un sujet du premier groupe[6].

$$S_{e \text{ groupe } 1} = 14\sqrt{1 - 0{,}75} = (14)\ (0{,}5) = 7$$

$$S_{e \text{ groupe } 2} = 7\sqrt{1 - 0{,}75} = (7)\ (0{,}5) = 3{,}5$$

Établissement de l'intervalle de confiance. Une fois l'erreur type connue, il est aisé d'établir un intervalle de confiance, sachant que la distribution des erreurs suit une courbe normale, symétrique de part et d'autre du score vrai et que sa moyenne est égale à zéro. En se référant aux proportions de la courbe normale (voir chapitre 7), on sait qu'il y a 68,3 % des chances que le score obtenu par une personne se situe entre plus ou moins une erreur type de son score vrai, 95,4 % entre plus ou moins deux erreurs types (ou $\pm$ 1,96 S_e si 95 %) et 99,7 % entre plus ou moins trois erreurs types (ou $\pm$ 2,58 S_e si 99 %).

5. Pour être plus précis, il s'agit en fait de ce que nous avons appelé plus tôt le résultat « systématique », afin de tenir compte de la présence possible d'erreurs systématiques.
6. Un autre problème est celui de savoir jusqu'à quel point l'erreur type de la mesure s'applique à tous les résultats obtenus à l'instrument, peu importe qu'ils soient faibles, moyens ou forts. Pour utiliser comme nous l'avons fait le concept d'erreur type, nous avons dû postuler que la distribution de l'erreur est relativement égale pour chacun des résultats obtenus : l'erreur type obtenue est une erreur moyenne, qui s'applique à tous les résultats. Toutefois, on peut avoir des raisons de croire que cela n'est pas toujours vrai, surtout lorsque certaines personnes répondent au hasard. Il existe des méthodes pour calculer l'erreur type de la mesure à un certain nombre de points ou d'intervalles le long de la distribution des résultats (voir Cronbach, 1990 ; *Standards*, 1999).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Ainsi, notre candidat contestataire avait raison : il est possible qu'il obtienne de 10 à 15 points de plus lors d'une nouvelle passation du test. La distribution des erreurs peut englober un tel écart, dont l'ampleur correspond à environ deux ou trois fois l'erreur type (soit $2 \times 6{,}3 = 12{,}6$ ou $3 \times 6{,}3 = 18{,}9$), alors que la distribution peut s'étendre sur une distance égale à six fois l'erreur type (lorsque trois erreurs types sont considérées de part et d'autre du score vrai).

Les erreurs de mesure se distribuent symétriquement de part et d'autre du score vrai, et non par rapport au score observé comme cela est faussement rapporté dans certains manuels (Nunnally et Bernstein, 1994). Donc, avant d'établir l'intervalle de confiance, il faut estimer le score vrai en utilisant la formule suivante :

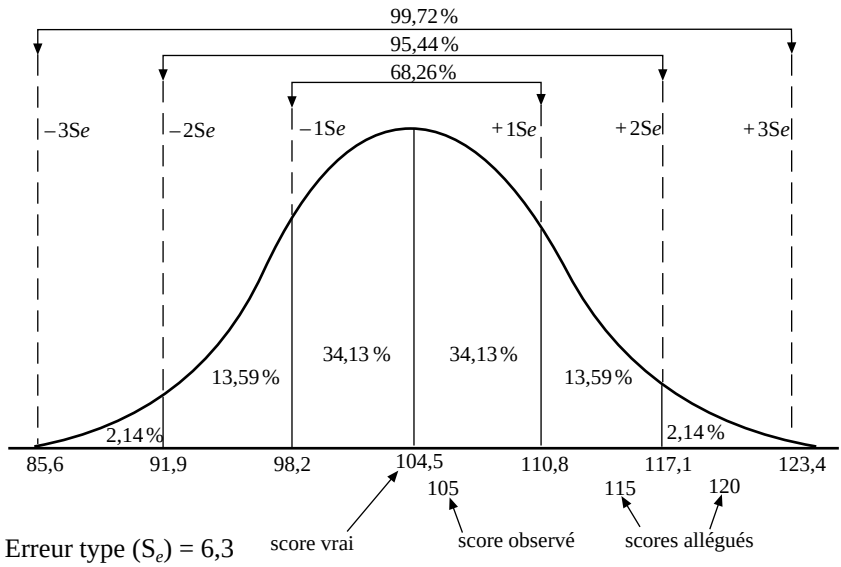
Résultat vrai (v) = $M_x + [(X - M_x) r_{xx}]$
où M_x : moyenne des résultats observés auprès de la population
X : résultat observé
r_{xx} : fidélité observée

Appliqué aux données de notre exemple et sachant que la moyenne de la population à ce test d'aptitude est de 100, le résultat vrai est estimé à 104,5 (soit 100 + [(105 – 100) 0,90]). Par définition, le score vrai se situe toujours entre le score observé et la moyenne[7]. Il est maintenant possible de représenter la distribution de l'erreur aléatoire et de déterminer l'intervalle de confiance (voir figure 4.2). Centrée sur le score vrai d'un individu, la distribution de l'erreur aléatoire de mesure s'étend de moins trois fois l'erreur type (soit $104{,}5 - (3 \times 6{,}3) = 85{,}6$) à plus trois fois (soit $104{,}5 + (3 \times 6{,}3) = 123{,}4$). Sans faire le calcul exact des probabilités, on peut voir dans cette figure qu'une différence de 10 à 15 points pour une même personne entre deux administrations du même test est possible. Ainsi, le candidat ayant obtenu 105 lors de sa première tentative peut obtenir 115 ou même 120 lors d'une deuxième tentative, selon que l'on retient un intervalle de confiance à plus ou moins deux (95,44 %) ou trois erreurs types (99,72 %).

7. Il existe une formule qui permet d'estimer l'intervalle de confiance pour cette estimation du score vrai (voir Nunnally et Bernstein, 1994).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 4.2
DISTRIBUTION DE L'ERREUR DE MESURE ALÉATOIRE ET INTERVALLES DE CONFIANCE AUTOUR DU SCORE VRAI



Usages de l'erreur type de la mesure. L'erreur type (S_e) sert à estimer la dispersion des résultats possibles pour une personne autour de son résultat vrai hypothétique. C'est un indice très utile en ce qui concerne le degré de précision d'un résultat, d'autant plus qu'il est exprimé dans les mêmes unités. En gestion des ressources humaines, les usages de l'erreur type sont nombreux (Gatewood et Feild, 1994 ; Guion, 1998). Premièrement, la connaissance de cet indice nous force à concevoir le résultat d'une personne, non pas comme un score précis, mais plutôt comme une **approximation** comportant une marge d'erreur ; le résultat obtenu par un individu à un moment donné ne représente qu'une possibilité parmi plusieurs résultats qu'il aurait pu avoir. Deuxièmement, l'erreur type est un **outil de décision** très précieux en ce qui concerne l'évaluation d'un candidat. Dans l'exemple présenté plus haut, le calcul de l'erreur type et des intervalles de confiance a permis de savoir qu'il était concevable qu'une personne augmente sa note de 10 à 15 points lors d'une éventuelle reprise. Sachant cela, l'organisation désirera peut-être mettre sur pied une politique autorisant les candidats à se présenter de nouveau au

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

test lors d'un concours subséquent. Par ailleurs, comme la marge d'erreur peut facilement être de six points, non seulement en moins mais aussi en plus, il se peut qu'un candidat atteigne la note de passage par chance. L'organisation pourrait alors exiger que tout candidat obtienne un résultat supérieur à une erreur type au-dessus de la note de passage, par exemple si ces employés opèrent une centrale nucléaire et que la sécurité des employés et du public ne peut être compromise. Troisièmement, l'erreur type peut servir à déterminer **l'écart minimal** pour que deux personnes soient considérées comme vraisemblablement différentes. Par exemple, un écart inférieur à une erreur type pourrait être jugé insuffisant, étant donné la grande probabilité qu'il soit le fruit du hasard.

APPRÉCIATION DU COEFFICIENT DE FIDÉLITÉ

Le coefficient de fidélité est un indice qui prend toute son importance au moment d'apprécier un instrument de mesure et ses résultats. Pour y arriver, plusieurs aspects sont à considérer : la méthode d'estimation employée, l'ampleur du coefficient calculé ainsi que certains facteurs pouvant influencer la valeur de ce coefficient. Et pour compléter l'analyse, rien de tel que d'examiner la signification concrète d'un coefficient en fonction des résultats obtenus et des marges d'erreur.

QUELLE EST LA MÉTHODE D'ESTIMATION EMPLOYÉE ?

La première question à se poser concerne le type de coefficient de fidélité, défini en fonction de la méthode d'estimation employée. Une section entière de ce chapitre a été consacrée à ces diverses méthodes et à leurs coefficients de fidélité qui ne sont pas interchangeables du point de vue des erreurs de mesure considérées. Or, il est essentiel, au point de départ, de *revoir attentivement la signification de chaque coefficient dans le contexte particulier de son application*. Par exemple, le coefficient de fidélité obtenu par **test-retest** peut être particulièrement significatif pour un usager qui se demande si les scores obtenus il y a un certain temps à cet instrument sont encore valables aujourd'hui. L'examen des moyennes et des écarts types entre les deux administrations peut aussi indiquer des changements survenus pendant l'intervalle de temps. L'estimation de la fidélité par **formes équivalentes** est une information importante si plusieurs versions du même

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

instrument sont envisagés. Quant au coefficient obtenu par bissection ou par indice alpha (α), il s'emploie pour estimer la **consistance interne** des tests et indique jusqu'à quel point les items sont homogènes. L'usager doit interpréter de tels indices avec prudence, en raison des nombreuses sources d'erreur qui ne sont pas vérifiées par ces méthodes ; les estimations de la fidélité en sont souvent surévaluées. Finalement, si le niveau de convergence entre examinateurs est une préoccupation, un coefficient de fidélité **interexaminateurs** fournira des indications appropriées. La comparaison des moyennes et des écarts types entre les examinateurs devra également être effectuée.

LE COEFFICIENT DE FIDÉLITÉ EST-IL SUFFISAMMENT ÉLEVÉ ?

Après avoir examiné la méthode d'estimation, il faut se poser la question suivante : est-ce que le coefficient de fidélité est suffisamment élevé ? Contrairement à ce que certains croient, il n'y a pas de seuil généralement approuvé par les experts, au-delà duquel un coefficient de fidélité est acceptable. Combien de fois n'a-t-on pas entendu le chiffre de 0,80 comme valeur minimale ? Malheureusement, les fondements scientifiques de cette assertion se font plus rares. On peut toujours affirmer qu'un coefficient doit être le plus élevé possible, mais cette affirmation ne tient pas compte de la réalité puisque la plupart des coefficients se situent entre 0,95 et 0,70, et parfois même à des niveaux aussi faibles que 0,50. La pratique étant l'art du possible, on doit alors considérer des aspects plus contextuels, comme l'usage qui est fait de l'instrument de mesure (Guion, 1993 ; Nunnally et Bernstein, 1994).

Une règle de conduite est suggérée. À cet égard, des spécialistes proposent la règle de conduite suivante : *plus une décision est importante, plus la fidélité des résultats sur lesquels s'appuie cette décision doit être élevée* (voir Gatewood et Feild, 1994). Autrement dit, il faut que le degré de fidélité soit proportionnel à l'importance des conséquences qui résultent de l'utilisation des résultats obtenus à un instrument de mesure. Par exemple, si un questionnaire est utilisé dans le cadre d'un sondage visant à connaître les attitudes des étudiants à l'égard de certaines méthodes pédagogiques, on peut tolérer que la fidélité de cet instrument ne soit pas très grande. En revanche, si un examen doit servir comme instrument de sélection pour l'admission à l'université, un coefficient de fidélité élevé est indispensable. On exigera

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

une fidélité encore plus grande si l'on se sert d'un test pour déterminer à quel moment une personne doit être placée dans une institution ou traitée pour un problème grave de comportement.

Encore une fois, il n'y a pas de normes précises pour évaluer l'importance des conséquences : cela dépend du point de vue où l'on se place. Parmi les nombreuses perspectives possibles en gestion des ressources humaines, deux sont singulièrement significatives. Il y a d'abord celle de l'organisation, de sa logique d'efficacité, de rentabilité, de climat de travail, d'intégrité, de développement, de survie, d'image corporative, de culture, etc. Il y aussi le point de vue de l'individu qui revendique l'accès à un revenu, à un environnement sécuritaire où il est agréable de travailler, à des tâches valorisantes, à un emploi qui accroît son sentiment de compétence et qui offre l'occasion de s'épanouir (un emploi qui n'est ni au-dessus, ni en dessous de ses capacités), et quoi d'autre.

Quelques valeurs repères. À la question du seuil minimal de fidélité acceptable, il arrive que des spécialistes se risquent à recommander des valeurs précises. Ainsi, pour des décisions importantes concernant des personnes, Nunnally et Bernstein (1994) soutiennent que 0,90 est un strict minimum et que 0,95 est un niveau plus désirable ; ce seuil de 0,90 est endossé par Cronbach (1990). Pour le contexte spécifique de la sélection du personnel, Gatewood et Feild (1994) trouvent acceptable un seuil de 0,85, mais préfèrent un coefficient de 0,90 et plus ; ces derniers rapportent l'opinion d'autres experts, notamment celle de Aiken (1988) pour qui le seuil devrait être de 0,85. Ces recommandations peuvent varier selon la méthode d'estimation utilisée (Womer, 1967, cité dans Gatewood et Feild, 1994). Avant de terminer, soulignons que tous les types d'instruments de mesure n'atteignent pas si facilement ces niveaux souhaités de fidélité. Si c'est le cas de la plupart des tests psychométriques d'aptitudes et de rendement les mieux conçus, il peut en être autrement des entrevues, des simulations et autres outils plus subjectifs. Malgré cela, il ne faut jamais rejeter un instrument uniquement parce qu'il est moins fidèle, sans au préalable avoir considéré la validité.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

QUELS SONT LES FACTEURS QUI PEUVENT INFLUENCER LA VALEUR DU COEFFICIENT ?

L'appréciation d'un coefficient de fidélité ne peut pas être complète sans analyser certains facteurs qui ont la faculté d'en influencer la valeur. Par exemple, un coefficient plus élevé n'est pas toujours garant d'une plus grande précision ; cela dépend du groupe de répondants. Décidément, la réalité n'est jamais aussi simple qu'on le voudrait.

Étendue des différences entre les répondants. Moins les répondants se distinguent par rapport à la dimension mesurée, moins le coefficient de fidélité est élevé. Si le groupe est relativement homogène, les répondants obtiennent des résultats passablement rapprochés ; la variance totale est alors petite. Cependant, il n'y a pas moins d'erreur aléatoire dans un échantillon homogène : l'erreur aléatoire est indépendante du résultat obtenu, qu'il soit faible, moyen ou élevé. Comme la fidélité est un ratio de l'erreur aléatoire (ou plutôt d'absence d'erreur aléatoire) sur la variance totale, par conséquent, si la variance totale est plus petite, la même quantité d'erreur apparaît plus élevée lorsqu'elle est exprimée en proportion de la variance totale. Cronbach (1990) donne l'exemple de deux groupes d'étudiants dont l'un, plus homogène, a un écart type de 9,1 pour ses résultats à un examen, comparativement à 10,8 pour l'autre groupe. Malgré la même quantité d'erreur aléatoire, la fidélité des résultats est estimée à 0,85 pour le premier groupe et à 0,90 pour le deuxième.

Ainsi, *un coefficient de fidélité moins élevé pour un groupe manifestant peu de variabilité dans le trait mesuré peut être aussi bon, même meilleur, qu'un coefficient plus élevé obtenu sur un groupe manifestant plus de variabilité*. Une façon de contrer cette distorsion est de calculer l'erreur type pour chaque sous-groupe afin de connaître l'étendue véritable de l'erreur. Cet indice tient compte de l'homogénéité des répondants, car la formule contient l'écart type des résultats (voir section portant sur l'erreur type). Pour l'exemple ci-dessus, les erreurs types sont 3,4 et 3,5 respectivement, ce qui est une différence négligeable.

Degré de difficulté de l'instrument par rapport aux répondants. La fidélité est plus élevée lorsque le degré de difficulté de l'instrument de mesure correspond au niveau d'habileté des répondants. Il y a trois raisons à cela. Premièrement, si un instrument est trop difficile ou hors de portée des répondants, ils seront plus enclins à **répondre au hasard**, n'ayant souvent pas d'autre choix ; on le sait, les réponses au

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

hasard diminuent la fidélité des résultats. Deuxièmement, lorsqu'un instrument est trop difficile ou trop facile, de nombreux répondants obtiennent sensiblement les mêmes résultats, soit très faibles, soit très forts ; cela peut avoir pour effet de **diminuer les différences entre les répondants**, ce qui, en retour, restreint le coefficient de fidélité (comme cela a été démontré au paragraphe précédent).

Par ailleurs, il est possible d'obtenir un coefficient de fidélité très élevé en conservant dans l'échantillon des personnes pour lesquelles l'instrument n'est pas approprié. On aurait un tel exemple si un examen était administré **à la fois** à des personnes ayant reçu des cours portant sur le sujet et à des personnes qui n'en auraient pas reçu. Certaines personnes obtiendraient alors des résultats très élevés et d'autres auraient pratiquement zéro. Il en résulterait une grande amplitude des résultats et un coefficient de fidélité singulièrement surévalué.

La troisième raison concerne les coefficients de fidélité obtenus par consistance interne, dont l'ampleur est reliée directement à la moyenne des **corrélations interitems**. Or, pour maximiser les corrélations interitems, les items doivent présenter à peu près la même difficulté, idéalement ni trop élevée, ni trop faible. Deux items mesurant le même trait mais dont le degré respectif de difficulté varie, auront une corrélation interitems plus faible. On peut aussi démontrer que les items d'un degré de difficulté moyen auront une variance maximale. Par exemple, un item réussi par la moitié des répondants et échoué par l'autre moitié sépare le groupe en deux sous-groupes égaux ; la dispersion est alors à son maximum par rapport à la moyenne de l'item ($p = 0{,}50$). Si les items mesurent le même trait, alors le fait de maximiser la variance amènera un coefficient de consistance interne accru.

En conclusion, *la difficulté d'un instrument doit être appropriée aux répondants pour lesquels on désire une précision maximale*. Ce principe ne veut pas dire que l'on doit tenir compte de tous les répondants. Par exemple, un examen de sélection servant à faire un premier filtrage des candidats ou à vérifier un niveau de maîtrise élémentaire doit pouvoir discriminer de façon fiable surtout au bas de l'échelle. Qu'il soit trop facile pour la majorité des répondants de sorte qu'il ne permet pas de distinguer les forts et les très forts est sans importance à ce stade-ci du processus. Réciproquement, si un examen est utilisé

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

pour repérer l'élite, il sera fatalement trop difficile pour la majorité. Enfin, il peut être préférable d'avoir des items comportant plusieurs degrés de difficulté afin de pouvoir discriminer les répondants visés tout au long de l'amplitude possible des scores.

Nombre d'items ou longueur de l'instrument. L'accroissement du nombre d'items ou l'augmentation de la période d'observation réduit les facteurs de chance. Un test de 30 items permet bien davantage aux sources d'erreur de s'annuler les unes les autres qu'un test de 10 items. De même, une période d'observation du comportement en groupe de 60 minutes risque d'être plus représentative qu'une période de 15 minutes. Plus un instrument est long, plus l'échantillon d'items ou d'observations a de chances d'être adéquat pour mesurer le trait en question. Cet aspect a déjà été abordé en début de chapitre (voir section « Le candidat », au paragraphe « 1. Tendance à répondre au hasard »).

L'effet de l'augmentation du nombre de mesures sur la diminution de l'erreur de mesure aléatoire est illustré par l'exemple suivant de Gatewood et Feild (1994). Un test qui comporterait 5 items et dont la fidélité serait estimée à 0,20, passerait à une fidélité de 0,33 si le nombre d'items était doublé. Si l'on poursuivait avec 20 items, puis avec 40 et 80 items, la fidélité s'accroîtrait respectivement à 0,50, à 0,67 puis à 0,80. On présume que les items ajoutés sont de même nature (contenu mesuré, degré de difficulté, etc.).

À sa manière, Cronbach (1990) illustre ce principe à l'aide du tableau suivant (voir tableau 4.7). Le tableau débute par le cas d'une mesure hypothétique dont la fidélité, estimée à 0,91, est déjà très élevée ; l'erreur type est de 1,00. Si l'on ajoute une deuxième mesure de même nature que la première, la fidélité passe à 0,95 et l'erreur type à 0,71, et ainsi de suite jusqu'à 100 mesures, où la fidélité est presque parfaite et l'erreur type, quasi nulle. Le tableau permet aussi d'analyser les effets sur les diverses composantes de la variance. La variance systématique demeure toujours la même ; ce qui change, c'est la diminution de la variance d'erreur et, évidemment, la variance observée puisque la variance d'erreur en fait partie. Au début, les gains en précision sont importants : la fidélité augmente rapidement alors que l'erreur type et la variance d'erreur diminuent tout aussi rapidement. Par la suite, il faut ajouter de plus en plus de mesures pour obtenir la même amélioration. Il faut noter cependant un phénomène

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

important : des gains très faibles de fidélité entraînent une diminution substantielle de l'erreur type, ce qui veut dire *qu'il vaut la peine d'améliorer la fidélité, même lorsqu'elle excède déjà 0,90*.

Tableau 4.7
EFFET DE L'AUGMENTATION DU NOMBRE DE MESURES SUR L'ERREUR ALÉATOIRE

Nombre de mesures	Variance			Coefficient de fidélité	Erreur type
	Erreur	Systématique	Observée		
1	1,00	9,80	10,80	0,91	1,00
2	0,50	9,80	10,30	0,95	0,71
3	0,33	9,80	10,13	0,97	0,58
5	0,20	9,80	10,00	0,98	0,45
10	0,10	9,80	9,90	0,99	0,32
100	0,01	9,80	9,81	0,999	0,10

Source : Tiré de Cronbach (1990), p. 207.

Une variante de la formule de Spearman-Brown (donnée plus tôt lorsque nous avons parlé de la méthode d'estimation par bissection) permet de déterminer le nombre de mesures (ou d'items) à ajouter à un instrument existant en vue d'accroître sa fidélité à un niveau désiré (Guion, 1998) :

$$N = \frac{r_{xx} \text{ désirée } (1 - r_{xx})}{r_{xx} (1 - r_{xx} \text{ désirée})}$$

où N : nouveau nombre d'items divisé par le nombre actuel

r_{xx} : fidélité de l'instrument actuel

r_{xx} désirée : fidélité désirée du nouvel instrument

Soulignons que N ne représente pas le nombre d'items à ajouter ou à soustraire, mais plutôt le nombre de fois qu'il faut multiplier le nombre actuel d'items pour obtenir le niveau de fidélité souhaité. Par exemple, à supposer que l'on désire augmenter à 0,90 la fidélité de 0,80 de l'instrument ci-dessus comprenant déjà 80 items, il

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

faudrait alors multiplier par 2,25 le nombre actuel d'items (soit [0,90 (1 – 0,80)] ÷ [0,80 (1– 0,95)]. De 80 items, l'instrument devrait passer à 180 (soit 2,25 × 80) !

Augmenter la fidélité peut nuire à la validité, ou dilemme étendue-précision. Augmenter le nombre d'items est une façon d'accroître la fidélité. Mais pour y arriver, il faut parfois sacrifier le nombre de dimensions évaluées et ainsi diminuer la validité : c'est le dilemme étendue-précision (Cronbach, 1990). Prenons l'exemple d'une entrevue de sélection. Dans un monde idéal, l'interviewer devrait disposer de tout le temps voulu pour évaluer l'ensemble des dimensions pertinentes avec autant de questions que nécessaire. Mais dans la réalité, l'interviewer dispose souvent d'une heure, de deux heures au plus. Il doit donc décider du nombre de dimensions à évaluer et du nombre de questions pour y parvenir. Un difficile arbitrage l'attend : opter entre la précision des informations recueillies ou leur étendue. S'il adopte la précision, l'interviewer choisit de se limiter à quelques dimensions, puis leur consacre toutes ses questions. Cette stratégie augmentera certes la fidélité des résultats obtenus, mais diminuera la validité des résultats en négligeant des dimensions pertinentes à l'emploi. Si l'interviewer préfère l'étendue, il élargit l'entrevue à toutes les dimensions pertinentes, quitte à poser peu de questions sur chacune ; ce qu'il gagne en étendue, il risque de le perdre en précision.

Un dilemme semblable se pose dans le choix des items lors de l'élaboration d'un instrument de mesure. Éliminer les items non homogènes, les items de l'instrument qui ne corrèlent pas avec les autres, est une autre manière de hausser la fidélité, lorsque la méthode d'estimation est basée sur la consistance interne. En effet, le calcul de l'indice alpha indique le degré de corrélation entre tous les items d'un instrument. Plus la corrélation entre les items est forte, c'est-à-dire plus les réponses des candidats à un item de l'instrument sont reliées à celles des autres items, plus l'indice alpha est élevé. Il est donc possible d'augmenter cet indice de fidélité en retirant les items qui ne corrèlent pas avec les autres au profit d'autres items plus homogènes. Mais si les items retirés mesurent des dimensions pertinentes au domaine de contenu, la validité de l'instrument risque d'en souffrir.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

QUE SIGNIFIE LE COEFFICIENT EN TERMES DE MARGE D'ERREUR ?

Afin de se doter de quelques valeurs repères, nous avons cité plusieurs experts en ce qui a trait au seuil minimal du coefficient de fidélité. Des indices de l'ordre de 0,85 à 0,90 ont alors été avancés. Pour le commun des mortels, sans doute habitués aux pourcentages et déjà convaincus que la perfection n'est pas de ce monde, ces chiffres semblent raisonnables. Quant aux spécialistes, la plupart ne semblent guère plus critiques : ils acceptent ces repères comme un dogme et les répètent à qui mieux mieux. Mais a-t-on vraiment essayé de voir ce qu'une telle précision veut dire, concrètement ? Est-ce qu'on en a évalué les répercussions sur l'organisation et sur les individus ? Voici quelques données réelles qui devraient nous donner matière à réflexion.

L'exemple d'un cas réel. Lors d'un processus de promotion interne dans un ministère, une épreuve du courrier a été utilisée comme outil de présélection pour des postes de supervision. Les candidats qui échouaient le test étaient automatiquement éliminés. Comme l'application de la grille de correction requiert un certain jugement de la part des correcteurs, le ministère, à la demande de candidats, a procédé à une deuxième correction. Quelques données sont compilées au tableau 4.8 pour 178 candidats. Entre les deux corrections, les résultats ont à peine augmenté de 0,2 points en moyenne, alors que le résultat le plus faible est demeuré à 6 et le plus fort est passé de 51 à 52 points. La corrélation entre les deux ensembles de résultats pour les 178 candidats est de 0,86, ce qui indique un niveau de fidélité acceptable selon les repères avancés plus tôt. Est-ce à dire qu'il n'y a pas de fluctuation d'une correction à l'autre pour un même candidat ? Certes non, car la variance d'erreur aléatoire est estimée à 16 %. Alors, de quelle ampleur sont ces fluctuations ? Les erreurs types, estimées autour de 3,0, laissent présager une marge d'erreur (S_e) pouvant atteindre facilement 6 ou même 9 points.

Jetons un coup d'œil aux différences de résultats réellement observées (troisième colonne). Bien que la différence moyenne soit minime, 0,2 point, ces différences varient d'une baisse de 13 points dans un cas à un gain de 16 points pour un autre, pour un test qui comporte 60 points et dont la moyenne est autour de 22 ! Pour avoir une idée plus exacte, la distribution de l'ensemble des différences est reproduite dans les deux dernières colonnes du tableau ; on constate que plusieurs candidats ont vu leur résultat changer considérablement.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 4.8
DISTRIBUTION DES RÉSULTATS OBTENUS PAR 178 CANDIDATS LORS DE DEUX CORRECTIONS SUCCESSIVES DE LA MÊME ÉPREUVE DU COURRIER (*IN-BASKET TEST*)

	Résultat obtenu			Distribution des différences de résultats	
	1re correction	2e correction	Différence	Classe	Fréquences
Moyenne	22,7	22,9	0,2	–1 3 à – 11	2
Écart type	7,8	8,2	4,2	– 10 à – 8	2
Minimum	6,0	6,0	–13,0	– 7 à – 5	14
Maximum	51,0	52,0	16,0	– 4 à – 2	42
				– 1 à + 1	58
				+ 2 à + 4	36
Corrélation (*r*)		0,86		+ 5 à + 7	15
(fidélité)				+ 8 à + 10	5
				+ 11 à + 13	2
Erreur type (S_e)	2,9	3,0	–	+ 14 à + 16	2
				Total	178

Par exemple, si l'on totalise le nombre de candidats qui ont subi des changements de huit points et plus, on obtient 13 personnes. Si l'on considère les changements de cinq points et plus, 42 personnes sur un total de 178 se retrouvent dans cette situation (soit 23,6 % de l'échantillon). Avec une note de passage fixée autour de 30 sur 60, de nombreux candidats qui avaient été acceptés lors de la première correction ne le sont plus, et inversement. La distribution des différences de résultats est illustrée à la figure 4,3[8].

Mot de la fin. À la lumière de tels résultats, on comprend mieux pourquoi 0,85 n'est pas un seuil de fidélité si rigoureux lorsque des décisions importantes sont prises concernant des individus. Il faut voir la marge d'erreur qui subsiste réellement dans les résultats, même

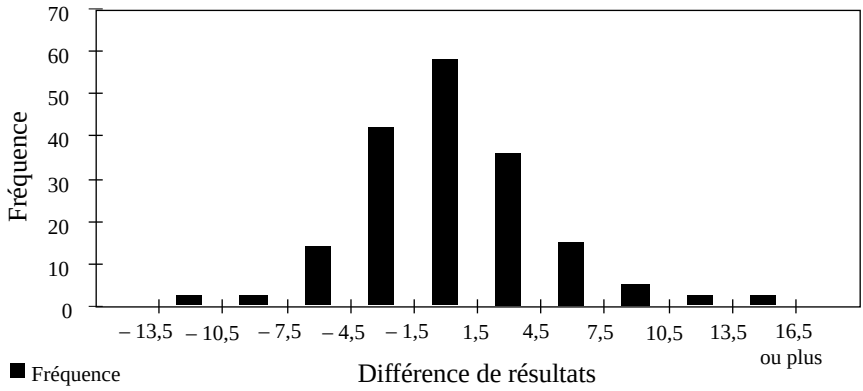
8. Voir Guion (1998) pour une autre illustration de la signification de l'erreur type.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

avec des indices de fidélité aussi élevés que 0,92 (voir Warner et Thissen, cité dans Guion, 1998), et encore là, il convient de ne pas confondre fidélité et validité. En effet, un instrument peut avoir une très grande fidélité et se révéler totalement invalide par rapport à l'objectif poursuivi.

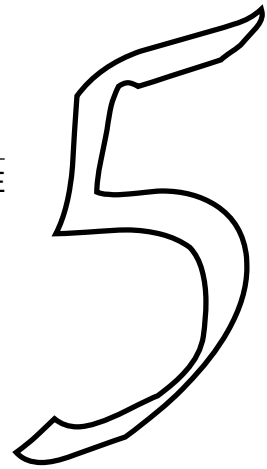
Figure 4.3
HISTOGRAMME DES DIFFÉRENCES DE RÉSULTATS APRÈS UNE SECONDE CORRECTION D'UNE ÉPREUVE DU COURRIER



© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

C H A P I T R E



ÉLABORATION D'INSTRUMENTS DE MESURE
Partie I : Détermination du domaine à mesurer

Élaborer un instrument de mesure, que ce soit un examen de connaissances, une entrevue structurée ou une grille d'évaluation du rendement, peut demander quelques heures, quelques jours ou plusieurs mois de travail à toute une équipe de spécialistes. Logiquement, les efforts consentis doivent être proportionnels aux conséquences en jeu. Par exemple, il est avantageux d'élaborer avec soin un instrument qui sera utilisé plusieurs fois, auprès de nombreux candidats, pour des engagements de longue durée ou qui peut être l'objet d'un litige. Par ailleurs, il y a des domaines qui, par nature, sont plus difficiles à mesurer que d'autres ; par exemple, il sera plus ardu d'évaluer l'ensembles des habiletés d'un bon chef d'équipe que de préparer un examen de connaissances portant sur l'usage d'un logiciel. Peu importe l'ampleur d'une telle démarche, il faut respecter certains principes pour obtenir un instrument rigoureux et efficace.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Évidemment, la validité de l'instrument de mesure sera au cœur de la démarche, car nous savons que, sans cette qualité première, un instrument ne peut être utile.

Ce chapitre porte sur l'élaboration d'instruments de mesure utilisés en gestion des ressources humaines. L'approche d'élaboration proposée s'appuie principalement sur la stratégie de validation basée sur le contenu[1]. Il y a plusieurs raisons au choix de cette approche. Nous avons vu que la validation basée sur le contenu est une approche rigoureuse, qui demande peu d'expertise technique en mesure et évaluation tout en offrant de bonnes garanties sur le plan de la défense des instruments en cas de litige. De plus, c'est l'approche qui convient le mieux à la plupart des situations de gestion des ressources humaines où les caractéristiques à mesurer sont constituées de connaissances, de compétences ou d'habiletés concrètes et observables directement dans l'emploi. L'approche est également praticable pour les caractéristiques personnelles plus abstraites, comme les traits de personnalité, la motivation ou les aptitudes générales, à condition que ces caractéristiques soient observables dans l'emploi (p. ex., garder son calme en présence de clients difficiles, persister dans l'action malgré les obstacles ou apprendre rapidement une nouvelle procédure).

Centré sur le contenu de l'emploi, le processus d'élaboration présenté dans ce chapitre est naturellement orienté vers des instruments de sélection[2] de type 1) examens de connaissances, 2) échantillons de travail ou *work sample tests* (test de dactylographie, de soudure, examen de conduite d'un véhicule lourd, etc. 3) mises en situation (épreuve du courrier ou *In-Basket Test*, exercice de prise de décision en groupe, etc. et 4) entrevues structurées suivant la forme situationnelle (p. ex., « que feriez-vous si vous vous trouviez dans telle situation ? ») ou comportementale (p. ex., « donnez un exemple où

1. Nunnally et Bernstein (1994) distinguent trois approches d'élaboration d'instruments, en fonction des trois stratégies traditionnelles de validation : validation basée sur le contenu, validation basée sur le construit et validation basée sur la relation avec d'autres variables (voir chapitre 2, section « Les stratégies traditionnelles de validation »).
2. L'expression « instruments de sélection » est une appellation générique courante qui désigne également les outils utilisés dans un processus de promotion ou de nomination interne.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

vous avez eu à traiter tel genre de problème »). Ce processus d'élaboration peut aussi servir à la conception d'instruments de mesure en évaluation du rendement et en formation du personnel.

Même en suivant une démarche relativement simple du point de vue technique, construire un instrument de mesure totalement nouveau peut exiger passablement de travail. Donc, avant de se lancer à corps perdu dans une telle entreprise, il faut vérifier s'il existe déjà un instrument de mesure qui correspond à nos besoins ou dont on pourrait au moins s'inspirer. Il est sans doute plus efficace d'adapter, voire d'améliorer un instrument déjà réalisé par des spécialistes que d'avoir à tout refaire soi-même. Les répertoires *Tests in Print* et *Mental Measurement Yearbook* publiés par The Buros Institute of Mental Measurement recensent plusieurs des tests disponibles sur le marché nord-américain[3]. On retrouve également dans la littérature des références à divers outils de mesure. En outre, des entités gouvernementales et de nombreuses organisations, surtout les grandes, ont développé leurs propres instruments de mesure ; on peut alors tenter d'obtenir une entente de collaboration. Les firmes de professionnels en gestion des ressources humaines ont aussi des instruments à offrir. Tous, cependant, ne disposent pas toujours de l'expertise pour élaborer des instruments dans les règles de l'art.

Plan du chapitre. Le contenu du chapitre est présenté dans quatre sections. La première offre une vue d'ensemble du processus général d'élaboration d'instruments de mesure fondés sur la validation de contenu. Les trois sections suivantes montrent en détail, étape par étape, comment définir clairement le domaine à mesurer par l'instrument. La suite du processus, soit le développement de l'instrument lui-même et son implantation, est exposée au chapitre suivant.

La personne qui n'est pas familière avec l'élaboration d'instruments de mesure trouvera la lecture des chapitres 5 et 6 difficile et même ennuyeuse. Si elle n'est pas déjà engagée dans l'élaboration d'un instrument, la démarche présentée lui paraîtra complexe et abstraite, tant il y a de détails, d'étapes à franchir, de principes à suivre. En revanche, pour une personne désireuse d'entreprendre l'élaboration d'un instrument, ces chapitres seront d'une aide précieuse tout au long de son travail.

3. http://www.unl.edu :80/buros/.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

PROCESSUS GÉNÉRAL D'ÉLABORATION D'INSTRUMENTS DE MESURE

Pour concevoir un instrument de mesure basé sur la validation de contenu, il faut définir le domaine que l'on veut mesurer, puis construire un instrument qui mesure adéquatement ce domaine (Cronbach, 1990 ; Nunnally et Bernstein, 1994). Le processus d'élaboration peut ainsi être divisé en deux phases, auxquelles une phase d'implantation a été ajoutée (voir tableau 5.1). La phase I consiste

Tableau 5.1
PROCESSUS GÉNÉRAL D'ÉLABORATION D'INSTRUMENTS DE MESURE FONDÉS SUR LA VALIDATION DE CONTENU

PHASE I.	**Finalités et spécification du domaine à mesurer**
Étape 1.	**Finalités de l'instrument de mesure :** préciser à quelles fins sera utilisé l'instrument de mesure (sélection, promotion, counselling d'emploi ou placement, formation, etc.).
Étape 2.	**Analyse et description de l'emploi :** relever les éléments observables, non triviaux et pertinents tels que ce qui doit être effectué (éléments, tâches et responsabilités), ce qui doit en résulter (produits ou résultats), le niveau de rendement attendu, le contexte de l'emploi, l'équipement et les technologies utilisés et, finalement, les exigences requises du point de vue de l'organisation.
Étape 3.	**Spécification du domaine ou du sous-domaine à mesurer :** choisir les éléments critiques et pertinents de l'emploi à être représentés dans l'instrument, et les transposer en termes 1) de comportements (éléments, tâches ou responsabilités), 2) de résultats produits par ces comportements ou 3) de connaissances, d'habiletés, d'aptitudes ou d'autres caractéristiques personnelles nécessaires à ces comportements.
PHASE II.	**Développement de l'instrument de mesure**
Étape 4.	**Conception de l'instrument dans son ensemble :** en fonction des ressources et des contraintes de l'organisation, préciser le format de l'instrument, le type d'items, la durée et le nombre approximatif d'items, le mode de correction, le processus d'interprétation et l'usage de normes, le nombre de versions de l'instrument, etc.
Étape 5.	**Création des items et élaboration des conditions d'application :** rédiger les items afin de représenter le domaine à mesurer (défini à l'étape 3) tout en respectant les spécifications de l'instrument désiré (précisées à l'étape 4). Rédiger les autres éléments de l'instrument comme les directives, le mode d'enregistrement des réponses, etc.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

à préciser les finalités de l'instrument de mesure (étape 1), à procéder à l'analyse et à la description de l'emploi (étape 2) pour finalement cerner le domaine à mesurer (étape 3). La phase II est consacrée à l'élaboration de l'instrument comme tel, de sa conception générale au prétest de la version expérimentale (étapes 4 à 8). La phase III complète le processus par la préparation de la documentation technique afférente (étape 9), l'implantation et le suivi périodique (étape 10).

Tableau 5.1 (suite)
PROCESSUS GÉNÉRAL D'ÉLABORATION D'INSTRUMENTS DE MESURE FONDÉS SUR LA VALIDATION DE CONTENU

Étape 6.	**Élaboration des outils d'évaluation :** préparer la clé de correction notamment en identifiant les éléments de réponse attendus, déterminer la façon de calculer les scores et, s'il y a lieu, les normes utilisées. Dans le cas d'un instrument qui requiert un processus d'observation, d'interprétation ou d'évaluation, élaborer les outils nécessaires à ces opérations.
Étape 7.	**Révision de la version expérimentale par des experts :** soumettre l'instrument de mesure à des experts afin d'examiner la clarté et la pertinence du contenu.
Étape 8.	**Essai de la version expérimentale et contrôle des qualités métrologiques :** faire passer l'instrument de mesure à un groupe témoin et, s'il y a lieu, modifier les directives, les items, les outils d'évaluation ou toute autre composante de l'instrument. Les données recueillies peuvent aussi servir à l'examen des qualités métriques et à l'établissement de normes et de notes de passage.
PHASE III. Implantation	
Étape 9.	**Rédaction des documents techniques :** rédiger un document qui décrit les étapes de la construction de l'instrument de mesure et un manuel de procédures qui indique notamment les usages prévus, la manière de l'administrer, de le corriger et d'en interpréter les scores.
Étape 10.	**Implantation et suivi :** planifier et organiser l'implantation de l'instrument, puis revoir périodiquement l'ensemble de l'instrument, ou lorsqu'il y a eu des changements dans la situation.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Assurément, un tel découpage en 10 étapes comporte une part d'arbitraire ; des étapes pourraient être fusionnées, alors que d'autres pourraient être subdivisées. Plusieurs auteurs proposent leur processus d'élaboration, chacun structuré de façon différente (Centre de psychologie du personnel, 1984 ; Gavin, 1977 ; Mussio et Smith, 1973 ; Nunnally et Bernstein, 1994 ; Plumlee, 1980 ; Schneider et Schmitt, 1986). L'important n'est pas le nombre d'étapes, mais l'ensemble des opérations à effectuer et leur séquence. Rappelons qu'il faut doser la rigueur et les efforts déployés à chacune des étapes selon les particularités de la situation et les conséquences envisagées (Nunnally et Bernstein, 1994).

ÉTAPE 1. FINALITÉS DE L'INSTRUMENT DE MESURE

Avant d'entreprendre tout travail d'élaboration d'un instrument de mesure, il faut s'interroger sur les objectifs poursuivis par les utilisateurs d'un tel outil. S'agit-il de sélection, de promotion de candidats déjà à l'emploi de l'organisation, de counselling d'emploi pour aider les candidats à s'orienter vers un emploi correspondant à leurs caractéristiques personnelles ou de l'établissement des besoins de formation chez des candidats ? Les objectifs conditionnent le contenu et la conception d'un instrument, et sans connaître le « pourquoi », il sera difficile de produire l'instrument efficace recherché.

Par exemple, un instrument devant servir à la **sélection** ou à la **promotion** du personnel devrait être à l'épreuve de la falsification des réponses de la part des candidats et facile à défendre en cas de contestation. De plus, le contenu devrait plutôt être orienté vers les caractéristiques générales et relativement stables de l'individu, car plus les caractéristiques sont profondes et génériques, plus elles seront difficiles à modifier par la suite, une fois la personne embauchée. Ainsi, les compétences évaluées seront constituées de connaissances et d'habiletés générales reliées aux exigences du poste plutôt que d'aspects très spécifiques à l'emploi et pouvant être appris en quelques semaines. Une attention spéciale devrait être accordée aux caractéristiques individuelles sous-jacentes au rendement et difficilement modifiables à court terme par la formation et l'entraînement.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

C'est tout le contraire pour un outil d'**évaluation du rendement** qui viserait à améliorer la performance et à diagnostiquer les besoins de formation. Il ne devrait pas porter sur les dimensions stables de l'individu, dont l'amélioration à court terme est plus ou moins sous son contrôle volontaire (p. ex., traits de personnalité ou aptitudes mentales générales). Il faut plutôt cibler des aspects modifiables de l'individu, comme ses connaissances et ses habiletés spécifiques à l'emploi, son comportement ou des éléments changeables de son rendement. De plus, un outil d'évaluation du rendement doit être particulièrement simple d'usage et peu exigeant pour les évaluateurs. S'il doit servir également à attribuer des primes et autres récompenses de ce genre, l'outil doit être conçu pour assurer l'équité entre les évalués et perçu comme tel. L'objectivité de la mesure devient un enjeu prioritaire, qu'il faut protéger autant des préjugés de l'évaluateur que des manipulations de l'évalué.

Le nombre et la diversité des emplois visés auront aussi des répercussions sur le degré de généralité du contenu à mesurer. Ainsi, un instrument de sélection qui ne doit servir que pour un emploi donné (p. ex., superviseurs des comptes à recevoir) sera plus spécifique qu'un autre conçu pour toute une famille d'emploi (p. ex., supervision de premier niveau). Le degré de généralité est un aspect des plus importants qui sera traité en profondeur à l'étape 3, lors de la spécification du domaine ou des sous-domaines à mesurer.

ÉTAPE 2. ANALYSE ET DESCRIPTION DE L'EMPLOI

L'analyse de l'emploi cherche à déterminer ce que la personne fait au travail, comment elle le fait, dans quel contexte et avec quelles ressources. L'analyse de l'emploi est le processus de collecte d'informations conduisant à la description de tous les aspects essentiels d'un emploi, qu'il s'agisse de tâches, de responsabilités ou de conditions de travail (Arvey et Faley, 1998 ; Harvey, 1991). Le but de l'analyse de l'emploi est d'arriver à circonscrire et à comprendre le mieux possible l'ensemble des éléments d'un poste de travail (Society of Industrial and Organizational Psychology, 1987) de façon à définir, à l'étape suivante, le domaine de contenu à mesurer. Il faut donc avoir une compréhension raisonnable des tâches à exécuter et de leur raison d'être si l'on veut connaître la performance désirée et préciser les compétences et les qualités ainsi requises chez le candidat (Guion, 1998).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

L'analyse d'emploi est le fondement du processus d'élaboration basé sur la validation de contenu. L'analyse d'emploi, souvent appelée analyse des tâches ou analyse de poste de travail, est primordiale en gestion des ressources humaines. Elle constitue la base à partir de laquelle les pratiques de gestion des ressources humaines sont élaborées (Spector, Brannick et Coovert, 1988). Il est impossible, par exemple, sans connaître le poste à combler et ses exigences, de développer des instruments de sélection, un système d'évaluation du rendement ou même un programme de formation adéquat. De plus, lorsqu'un instrument de mesure est élaboré avec la démarche de la validation de contenu, il faut que cet instrument reflète le contenu de l'emploi. À la base de la démarche, il doit donc y avoir une analyse d'emploi fiable et complète, comme en font foi les lignes directrices prescrites par les grands organismes professionnels (SCP, 1987 ; SIOP, 1987 ; *Standards*, 1999) et maintes fois confirmées par les tribunaux (Arvey et Faley, 1988 ; Mussio et Smith, 1973 ; Thompson et Thompson, 1982).

Une démarche d'analyse de l'emploi spécifique à l'élaboration d'instrument. L'analyse de l'emploi est un domaine fort élaboré de la psychologie industrielle et de la gestion des ressources humaines. Comme nous ne pouvons rendre compte ici de l'ensemble des connaissances et des techniques utilisées en ce domaine, nous invitons le lecteur intéressé à consulter les nombreux ouvrages qui y sont consacrés[4]. Toutefois, étant donné le rôle majeur de l'analyse de l'emploi dans l'élaboration et la défense d'un instrument de mesure fondé sur la validation de contenu, il est essentiel d'exposer une démarche suffisamment complète pour être utilisée en pratique. La méthode d'analyse d'emploi présentée est spécifique à l'élaboration d'instruments de mesure. Cette approche ne pourrait pas servir, par exemple, à établir une classification salariale ou à déterminer une ligne de progression.

4. Voir notamment les travaux de Gatewood et Feild (1994), de Schmitt et Chan (1998) ainsi que de Goldstein, Zedeck et Schneider (1993) portant sur l'analyse de l'emploi orientée vers la sélection du personnel ou l'élaboration d'instruments de mesure.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Une démarche qui sert d'abord à décrire l'emploi. Harvey (1991) rappelle avec insistance, et avec de nombreuses citations à l'appui, que l'analyse de l'emploi ne devrait essentiellement servir qu'à décrire l'emploi à partir de ses **aspects observables** et vérifiables. Elle devrait exclure le processus d'identification des exigences de l'emploi (*job specifications*, *worker specifications*), qui consiste à inférer les aptitudes et autres caractéristiques individuelles requises chez les candidats (capacité d'apprentissage, coordination motrice, leadership, ténacité, etc.), parce que c'est un processus distinct qui dépasse la simple description de l'emploi lui-même et de son contexte ; Grant (1989) partage cet avis. L'établissement des exigences sera accomplie, mais seulement à l'étape suivante, lors de la spécification du domaine à mesurer (voir étape 3).

ASPECTS DE L'EMPLOI À CONSIDÉRER

L'analyse de l'emploi n'a pas à inclure tout ce qui se passe au travail ; il n'est pas nécessaire qu'elle contienne les aspects triviaux ou sans importance de l'emploi (Arvey et Faley, 1988). Que les fauteuils soient gris ou que le stationnement soit à gauche de l'édifice ne devraient pas être des éléments à retenir. Il n'est pas indispensable non plus que la description porte sur tout l'emploi. Dans certaines circonstances, il serait possible de limiter la description de l'emploi aux éléments pertinents au domaine de contenu envisagé. Par exemple, dans le cadre de l'élaboration d'un examen sur l'entretien quotidien d'une photocopieuse pour des préposés à la reprographie, il n'est sans doute pas très important de décrire en détail les aspects de l'emploi concernant les relations avec la clientèle.

Le tableau 5.2 présente les principaux aspects pouvant être considérés lors de l'analyse d'emploi : 1) les informations concernant l'identification de l'emploi, 2) sa raison d'être, 3) ce qui doit être effectué, 4) ce qui doit en résulter, 5) les normes formelles de rendement, 6) le contexte de l'emploi et 7) les équipements utilisés et les technologies requises (Tziner, Jeanrie et Cusson, 1993). À ces aspects observables, on peut ajouter 8) les qualifications et les exigences requises telles qu'elles sont perçues par l'organisation ou les autres partenaires. Ensemble, ces aspects regroupent les éléments proposés par Guion (1998) à partir de la méthode du United States Employment Service (USES).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 5.2
ASPECTS POUVANT FAIRE PARTIE DE L'ANALYSE ET DESCRIPTION DE L'EMPLOI

- **Identification de l'emploi**
  - titre de l'emploi, classification, etc.,
  - unité administrative d'appartenance,
  - supérieur hiérarchique.
- **Sommaire de l'emploi ou sa raison d'être**
- **Ce qui doit être effectué**
  - éléments (*job elements*),
  - tâches (*tasks*),
  - Responsabilités (*responsibilities*, *activities* ou *duties*) ou fonctions (*functions*).
- **Ce qui doit en résulter**
  - produits ou résultats*,
  - conséquences de l'emploi sur la réalisation de la mission et des objectifs, le déroulement des programmes ou autres.
- **Normes formelles de rendement ou critères d'évaluation du rendement**
  - normes qualitatives de rendement,
  - normes quantitatives de rendement.
- **Contexte de l'emploi**
  - environnement physique : espace, température, propreté, bruit, et autres particularités,
  - conditions de travail : horaire, rémunération, contrat de travail, etc.,
  - environnement administratif : mission, structure et processus de coordination, niveau de responsabilité, relations avec les autres emplois, etc.,
  - environnement psychologique et social : culture et climat.
- **Équipement et technologies**
  - équipement, outils, appareils, etc.,
  - connaissances spécialisées, technologies, techniques, etc.
- **Exigences requises (du point de vue de l'organisation)**
  - scolarité et formation,
  - expérience,
  - connaissances (générales et spécifiques),
  - aptitudes et habiletés,
  - autres caractéristiques personnelles.

* Il se peut que les produits ou les résultats soient déjà clairement indiqués dans « Ce qui doit être effectué » ou implicites et qu'il ne soit donc pas nécessaire de les préciser davantage.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Identification et sommaire de l'emploi. Avant d'entreprendre l'analyse des principaux aspects de l'emploi, il convient de rassembler les informations servant à identifier l'emploi, comme le titre de l'emploi, sa classification s'il y a lieu, l'unité administrative à laquelle il est rattaché, le titre de l'emploi du supérieur immédiat et toute autre information pertinente. Ensuite, il est important de saisir l'ensemble de tout ce qui compose l'emploi avant d'en examiner les parties. Pour ce faire, il est intéressant de dresser un **sommaire de l'emploi** à travers ses fonctions ou responsabilités principales. Connaître la **raison d'être de l'emploi** ou les objectifs justifiant son existence est une autre façon d'en saisir l'essence. Après cette étape, on devrait pouvoir comprendre la relation entre l'emploi et la mission générale de l'organisation ainsi que les principales responsabilités de l'unité administrative concernée (décrites plus loin à la section « Contexte de l'emploi », sous l'élément « Environnement administratif »).

Ce qui doit être effectué. L'analyse de l'emploi est ensuite dirigée sur « ce qui doit être effectué ». C'est la partie la plus abondante de l'analyse ; les approches présentées dans les manuels portent souvent sur ce seul aspect. L'analyse peut se faire à divers niveaux, selon la taille des unités d'analyse considérées. Notons qu'il n'y a pas de véritable consensus sur ces **niveaux d'analyse** ni sur leur définition. Néanmoins, selon plusieurs spécialistes (Guion, 1998 ; Tziner, Jeanrie et Cusson, 1993), on peut définir les trois niveaux suivants : l'élément, la tâche et la responsabilité (voir tableau 5.3). Les étiquettes attribuées à chacun de ces niveaux et peuvent varier d'un auteur à l'autre ; celles que nous utiliserons proviennent de Guion (1998).

Au premier niveau se situe **l'élément** (*job element*) qui constitue la composante la plus simple de l'emploi ; l'emploi peut difficilement être décomposé en entités plus petites. Elle apparaît sous forme de paroles ou d'actions peu complexes : « desserrer un boulon », « mesurer l'usure d'un piston », « écrire une adresse » et « répondre au téléphone » en sont des exemples.

L'ensemble des éléments appartenant à une étape d'un processus de travail forme une **tâche** (*task*), soit le deuxième niveau d'unités d'analyse. Ainsi pour un emploi de cadre, « la préparation des budgets » pourrait être une tâche composée d'éléments comme « consulter les membres du service », « analyser les budgets précédents » ou « établir les priorités ». Une autre des tâches connexes à la gestion du budget

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 5.3
NIVEAUX D'UNITÉS D'ANALYSE

Niveau d'unités d'analyse	Définition	Exemples
Élément (*job element*)	Composante la plus simple. Paroles ou actions peu complexes.	Desserrer un boulon. Mesurer l'usure d'un piston. Inscrire une adresse. Répondre au téléphone.
Tâche (*task*)	Ensemble des éléments appartenant à une étape d'un processus de travail, comportant un début et une fin clairement identifiables.	Préparation du budget. Contrôle des dépenses. Tri des lettres auxquelles on doit répondre immédiatement.
Responsabilité (*responsibility, activity* ou *duty*)	Ensemble des tâches appartenant à un processus complet de travail, appelées aussi fonctions.	Gestion budgétaire Traitement du courrier Service à la clientèle Supervision du personnel Relations publiques Établissement des objectifs Communication
Fonction (*function*)	Responsabilité appartenant à une typologie très répandue en gestion.	Planification Organisation Direction Contrôle

pourrait être le « contrôle des dépenses », comportant les éléments suivants : « approuver des dépenses », « enregistrer des dépenses aux divers postes budgétaires » et « modifier le budget en cours d'opération s'il y a lieu ». Une tâche a un début et une fin clairement identifiables.

Lorsque sont réunies les diverses tâches d'un processus complet de travail, elles forment une responsabilité (*responsibility*, *activity* ou *duty*), qui représente le troisième niveau d'unités d'analyse. Ainsi, les diverses tâches relatives au budget pourraient constituer la **responsabilité** appelée « la gestion budgétaire ». Pour un emploi de cadre, d'autres exemples de responsabilités pourraient être « le service à la clientèle », « la supervision du personnel », « les relations publiques », « l'établissement des objectifs du service ». Les responsabilités correspondent plus ou moins aux diverses raisons d'être de l'emploi. Les activités prennent aussi parfois le nom de fonctions. En gestion, les fonctions

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

renvoient généralement à la planification, à l'organisation, à la direction (qui inclut notamment la communication, la mobilisation et le coaching) et au contrôle[5].

La distinction entre ces trois niveaux d'unités est plus relative qu'absolue ; la démarcation entre chacun d'eux dépend des regroupements que l'analyste voudra bien faire avec les divers éléments de l'emploi qu'il aura conservés. Il n'est pas nécessaire d'avoir recours aux trois niveaux d'analyse ; selon les circonstances, on peut retenir un, deux ou trois niveaux. Cependant, dans la plupart des cas en contexte de sélection ou de promotion du personnel, lorsque l'emploi est le moindrement complexe, l'analyse porte sur le niveau des « tâches », alors que le niveau « responsabilités » ou « fonctions » sert habituellement à regrouper ces tâches en sous-ensembles. Il n'est pas souhaitable que l'analyse ne retienne que le niveau « responsabilités » ou « fonctions », ce niveau n'étant pas assez spécifique.

Le niveau d'analyse choisi, il faut relever et décrire chacune des tâches (ou éléments). Afin d'avoir toutes les informations requises, il faut suivre quelques règles d'usage (Gatewood et Feild, 1998 ; Goldstein, Zedeck et Schneider, 1993). Chaque tâche doit être décrite au moyen d'une **proposition**, suffisamment claire et spécifique, qui répond à quatre questions : 1) que fait l'employé (quoi), 2) comment le fait-il, 3) à qui ou à quoi le fait-il et 4) pourquoi. De plus, une proposition doit commencer par un verbe d'action. Deux exemples sont présentés dans la figure suivante (voir figure 5.1).

Les tâches ou les éléments analysés n'ont pas tous la même importance ; il faut donc préciser la **fréquence** (ou durée) de chacun et leur **importance** (*Standards*, 1999, article 14.10). Une façon d'obtenir des informations permettant d'apprécier l'importance des composantes et leur degré de difficulté est de s'enquérir des principaux problèmes rencontrés et de leurs répercussions sur le travail, sur celui des autres employés ou sur tout autre aspect de la réalisation de la mission de l'organisation.

---

5. On retrouve parfois des descriptions de tâches qui comportent une rubrique consacrée aux communications : communications internes et externes à l'organisation, personnes impliquées, moyens utilisés, fréquence, etc. De la même manière, on peut inscrire dans une rubrique à part les données concernant le personnel sous la supervision du titulaire de l'emploi : nombre de personnes, leur emploi, leur classement, etc.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 5.1
EXEMPLES DE PROPOSITION POUR DÉFINIR LES TÂCHES D'UN EMPLOI

Une proposition doit décrire ce que fait l'employé (quoi), comment il le fait, à qui ou à quoi il le fait et pourquoi.

Exemple pour un poste de conseiller en gestion des ressources humaines.

Quoi? Dresse la liste
À qui/À quoi (objet)? des sources de recrutement
Comment? en fonction de leurs coûts et de leur efficacité
Pourquoi? pour les présenter au gestionnaire (client interne) et entreprendre le recrutement.

Exemple pour un poste de secrétaire*.

Quoi? Trie
À qui/À quoi (objet)? la correspondance
Comment? par ordre alphabétique
Pourquoi? pour en faciliter le classement.

* Exemple tiré de Schmit et Chan (1998), p. 45.

Il est fortement recommandé de **regrouper** les tâches ou les éléments en grandes catégories (p. ex., classer les tâches en quatre à sept grandes responsabilités ou fonctions) afin de structurer davantage l'information et de faciliter la compréhension de ce qui doit être effectué dans l'emploi. Par exemple, les tâches d'un directeur municipal ont été regroupées pour former les six grandes responsabilités suivantes : 1) la gestion des budgets, 2) la gestion de la réglementation, 3) la gestion des plaintes et des réclamations, 4) la gestion des programmes et des projets, 5) la gestion interne et la supervision du personnel et 6) la gestion des communications (voir figure 2.1). De plus, à l'intérieur de chaque regroupement, il peut être utile de classer les éléments ou les tâches en suivant une certaine logique, par exemple en commençant par les plus importants ou selon un ordre chronologique d'exécution.

Ce qui doit en résulter. L'analyse de l'emploi ne doit pas seulement prendre en considération « ce qui doit être effectué » dans le poste en question, mais aussi ce qui doit résulter de chaque élément, tâche ou responsabilité, selon le niveau d'analyse choisi ; on parle alors des **produits** ou des **résultats**. Par exemple, souder un tuyau devrait donner lieu à une soudure qui résiste à telle pression ; mesurer l'usure d'un piston doit permettre au mécanicien de connaître avec

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

précision l'usure du piston par rapport aux normes acceptées par le manufacturier ; la préparation des budgets doit résulter en des budgets rédigés selon les règles et les procédures en vigueur dans l'organisation.

Souvent, les propositions décrivant les tâches (rédigées à l'étape précédente) incluent déjà dans le « comment » ou dans le « pourquoi » ces résultats et ces produits. Par exemple, « Préparer les états financiers selon les règles et procédures en vigueur dans l'organisation », « Rencontrer périodiquement les clients afin de mieux en connaître leurs besoins » ou « Faire la promotion de nos services pour augmenter le taux d'utilisation de nos ressources ». Dans pareil cas, il n'est pas nécessaire de répéter « ce qui doit en résulter » ; cette rubrique doit simplement être omise de la description de l'emploi. Il peut arriver aussi que le résultat ou le produit soit implicite dans la proposition. Par exemple, « Rédiger les procès-verbaux », « Prendre les appels » ou « Répartir le travail équitablement entre les employés ».

Normes formelles de rendement. Pour comprendre toutes les exigences d'un emploi, il faut connaître en outre le rendement attendu par l'organisation. Les normes formelles de rendement peuvent être définies en termes **quantitatifs** (nombre d'unités produites, fréquence des erreurs commises, etc.) ou **qualitatifs** (participe activement aux réunions, est ouvert aux nouvelles technologies, respecte strictement les règles établies, etc.). Lorsqu'il existe un système **d'évaluation du rendement**, il s'agit simplement de s'informer des critères servant à évaluer la performance des employés. Sinon, on peut découvrir ces normes en cherchant à connaître les employés qui ont obtenu les promotions, ceux qui sont reconnus comme les employés « modèles », ceux qui sont dans les bonnes grâces du supérieur, et chercher à comprendre pourquoi il en est ainsi. Inversement, on peut s'informer sur les employés qui ont été remerciés de leurs services ou placés hors d'état de nuire, les employés qui ont été marginalisés par l'organisation, etc.

Une démarche complète d'analyse de tâche verrait à identifier les normes de rendement telles qu'elles sont spécifiées par les **divers constituants** de l'organisation. L'expression « constituant », empruntée de Tsui (1984, 1987), désigne toutes les personnes, les institutions ou les autres interlocuteurs qui transigent avec un service ou un employé dans une organisation. Comme chaque constituant a des attentes qui lui sont propres, il existe autant de barèmes pour estimer la performance

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

du service ou de l'employé. En plus des supérieurs, qui sont les constituants habituellement les plus importants, il y a aussi les collègues, les subordonnés, ainsi que les clients internes et externes à l'organisation.

Contexte de l'emploi. Le contexte de l'emploi doit faire partie de l'analyse. Les principaux éléments à considérer sont, bien entendu, l'**environnement physique** (espace, température, propreté, bruit, etc.) et les **conditions de travail** (horaire, rémunération, contrat de travail, etc.). Vient ensuite l'**environnement administratif**, composé de divers éléments comme la *mission* de l'organisation, sa raison d'être et les grandes responsabilités qui en découlent pour l'unité concernée. Il y a aussi la *structure*, qui prend la forme d'organigramme, et les *processus de coordination*. Sous ce rapport, il est particulièrement capital de circonscrire le *niveau de responsabilité* ou la marge discrétionnaire laissée au titulaire de l'emploi. Par exemple, quelles sont les décisions qui peuvent être prises ou les activités qui peuvent être réalisées sans l'intervention préalable, la vérification ou l'approbation d'une autorité supérieure ? Réciproquement, quelles sont les décisions ou les activités du titulaire qui doivent être vérifiées par ses supérieurs ou pour lesquelles il doit les consulter ou obtenir leur approbation ? Il est également important de cerner les *relations avec les autres emplois* dans l'organisation.

Enfin, il y a l'**environnement psychologique et social** de l'emploi, dont les principaux aspects sont la culture et le climat organisationnels. La *culture* se rapporte aux valeurs partagées par les différents groupes au sein de l'organisation, c'est-à-dire leurs façons habituelles de penser et d'agir (Maillet, 1993 ; Ouchi et Wilkins, 1985). Par exemple, le respect de la clientèle, le meilleur produit au moindre coût, la forte décentralisation des prises de décision ou une norme informelle de rendement élevé pourraient être des éléments de culture dans une organisation. Quant au *climat*, il se définit par toutes les caractéristiques psychosociales de l'organisation telles qu'elles sont perçues par ses membres. Il comprend les comportements des individus, leurs statuts et leurs rôles, la dynamique des groupes, les systèmes d'influence, l'exercice du pouvoir, etc. (Chagnon, 1991).

Équipements et technologies. L'analyse de l'emploi doit inclure les équipements et les technologies à maîtriser dans l'emploi pour réaliser les produits et les services. Par exemple, pour occuper un emploi de

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

secrétaire, la personne doit connaître le fonctionnement d'équipements tel un micro-ordinateur ou un télécopieur et des technologies comme un logiciel de traitement de texte ou la procédure d'appel conférence, etc.

Exigences requises. Les exigences requises portent sur les compétences et les autres caractéristiques personnelles que les candidats doivent posséder pour réussir dans cet emploi. On retrouve d'abord les qualifications de base comme la scolarité et la formation spécialisée, l'expérience de travail et les connaissances. Ensuite, il y a les autres aptitudes, habiletés et caractéristiques personnelles, comme la capacité d'analyse, le jugement, la persévérance ou le sens des responsabilités.

Même si ces exigences ont déjà été déterminées et figurent dans la description de l'emploi préparée par l'organisation, elles doivent être étudiées de nouveau, en interrogeant, comme pour les normes de rendement, les **principaux constituants** de l'organisation, ou du moins les titulaires de l'emploi concerné et les supérieurs immédiats. Il peut être utile de leur poser des questions du genre suivant : « Selon vous, quelles sont les qualités nécessaires pour avoir un bon rendement à ce poste ? » ; « Selon vous, qu'est-ce qui distingue ceux qui ont le plus de succès à ce poste ? »

Il ne s'agit **pas** à ce stade-ci de tenter d'**inférer les exigences requises** par l'emploi, mais simplement de noter ce que les principaux intéressés en pensent. Ce n'est qu'à l'étape suivante que l'analyste essayera d'inférer les exigences devant être mesurées par l'instrument de mesure. Les opinions des principaux intéressés recueillies lors de la présente étape à ce sujet pourront alors compléter l'ensemble des informations rassemblées depuis le début du processus d'analyse d'emploi.

MÉTHODOLOGIE

Modalités et techniques. Pour recueillir les informations sur les divers aspects de l'emploi, il existe plusieurs **modalités**, notamment l'observation, les entrevues individuelles, les entrevues de groupe, les questionnaires ou le journal de travail. Évidemment, l'observation n'est utilisable que pour les aspects visibles de l'emploi alors que les autres modalités, basées essentiellement sur le questionnement, peuvent s'appliquer à tous les aspects. Il est également utile de consulter les descriptions d'emploi existantes de même que les autres documents pertinents comme les manuels de formation, la Classification nationale descriptive des professions (Développement des ressources humaines du Canada, 1993, 1995). S'il s'agit d'un emploi syndiqué,

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

il faut consulter la convention collective. En plus de ces modalités, diverses **techniques** ont été élaborées par les auteurs, telles que la méthode des incidents critiques (Flanagan, 1954) et ses variantes (McClelland, 1976 ; Spencer et Spencer, 1993) ainsi que de nombreuses méthodologies quantitatives (Gatewood et Feild, 1998 ; Guion, 1998 ; Tziner, Jeanrie et Cusson, 1993).

Experts participants (« subject matter experts » ou SME). Quelle que soit la méthodologie choisie, les informations devront être recueillies auprès de plusieurs personnes, principalement des titulaires de l'emploi et des supérieurs (Goldstein, Zedeck et Schneider, 1993 ; Schmitt et Chan, 1998). Ces personnes doivent bien connaître l'emploi. Les titulaires doivent être en poste depuis suffisamment longtemps pour avoir accompli toutes les tâches dans divers contextes. À noter que les supérieurs sont plus utiles s'ils participent encore activement aux tâches de l'emploi que ceux qui ne font que jouer un rôle administratif ou sont trop éloignés des opérations. Les titulaires ont plus de facilité à décrire exactement leurs tâches, le contexte dans lequel ils travaillent ainsi que l'équipement et les technologies utilisées, alors que les superviseurs sont censés connaître la raison d'être de l'emploi, ce qui doit en résulter, les normes de rendement et les exigences requises.

Pour certains types d'informations, par exemple le rendement attendu ou les exigences requises, il peut être utile de consulter d'autres constituants tels que des **subordonnés**, des **clients** ou toute personne pouvant fournir des informations pertinentes. Toutes ces personnes constituent des experts de l'emploi visé ou des *subject matter experts* (SME) ; elles doivent être clairement qualifiées pour ce rôle et avoir reçu une formation en ce sens (SIOP, 1987).

Combien de personnes devront ainsi participer à titre d'experts (SME) ? Un **échantillon représentatif** (Arvey et Faley, 1988) **de trois à six personnes** (Schmitt et Chan, 1998) peut suffire. Il faudra toutefois veiller à ce que le groupe ne manifeste aucun préjugé par rapport aux principaux paramètres comme le sexe, la race, l'origine ethnique, etc., de manière à ne pas fausser les résultats. Par exemple, Landy et Vasey (1991) ont observé que l'ancienneté des titulaires d'emploi influençait leur estimation de la fréquence de certaines tâches. Il ne faut pas oublier que les **responsables de l'analyse** doivent aussi avoir les compétences et l'intégrité requises (Arvey et

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Faley, 1988). À cet égard, il serait plus sage de recourir à plus d'un analyste afin de diminuer les risques reliés aux préjugés et les erreurs aléatoires (voir chapitre 4).

Rédiger un document qui décrit les divers aspects de l'emploi et un rapport qui relate la démarche suivie. Une fois les renseignements recueillis et analysés, un document servant à décrire l'**emploi** devra être rédigé. Il ne s'agit pas d'élaborer une description d'emploi traditionnelle comme on en retrouve dans les organisations et qui porte surtout sur les tâches et responsabilités. On doit plutôt rassembler tous les renseignements recueillis de façon cohérente dans le but de préciser, à l'étape suivante, les exigences requises par cet emploi et qui formeront le domaine de contenu à mesurer. L'ensemble des informations peut être organisé en fonction des divers aspects couverts par l'analyse (voir tableau 5.2). Les tâches ou les éléments décrits au moyen de propositions à la rubrique « Ce qui doit être évalué » demeurent l'aspect le plus important de la description de l'emploi. Les autres aspects complètent cette information, en permettant notamment de mieux comprendre les tâches, leur raison d'être et leur importance ; ils aideront éventuellement à identifier les exigences requises pour effectuer les tâches de l'emploi.

Pour plus de sûreté, il faut vérifier que les renseignements ont été compris, formulés et organisés correctement par le responsable de l'analyse. Cette vérification est effectuée généralement en soumettant le document aux détenteurs d'emplois, à leurs supérieurs immédiats ou à tout autre expert de l'emploi (SME).

Enfin, il importe de constituer un rapport qui relate en détail l'ensemble de la **démarche** suivie ; cela est devenu une exigence dans plusieurs décisions de la cour américaine (Thompson et Thompson, 1982). Un tel rapport pourrait comporter, en plus du document plus haut, les rubriques suivantes : 1) la description des techniques de collecte d'informations utilisées, 2) la documentation analysée (descriptions de tâches, manuels de formation, etc.), 3) les titulaires d'emplois et les autres personnes ayant fourni les informations (titre, niveau hiérarchique, qualifications, taille de l'échantillon), et 4) les qualifications des personnes ayant réalisé l'analyse (Gavin, 1977).

En cas de **litige** portant sur la validité d'une méthode de sélection, la qualité du processus d'analyse et de description de l'emploi pourra être déterminante et l'application des recommandations et

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

principes mentionnés tout au long de cette section deviendra un enjeu important. Plusieurs de ces aspects ont aussi été confirmés par les tribunaux américains (Thompson et Thompson, 1982) : 1) des modalités et des techniques de collectes de données systématiques et objectives, 2) bien documentées, 3) qui donnent lieu à des informations à jour et représentatives de l'emploi, 4) permettant d'en identifier les aspects importants et critiques, 5) et cela avant d'inférer les exigences requises par l'emploi pour ce qui est des connaissances, des habiletés ou des aptitudes.

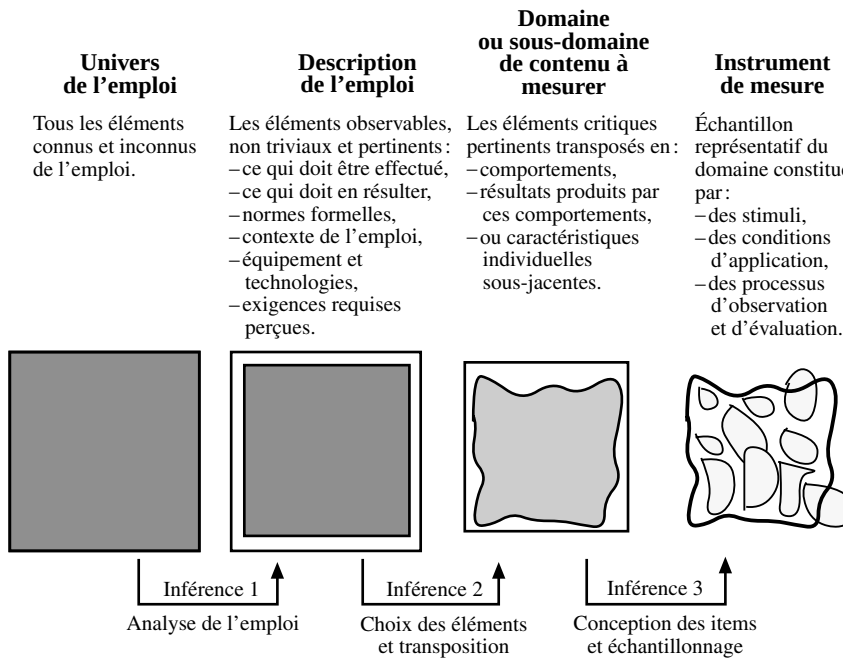
PROCESSUS DE DÉFINITION DU DOMAINE DE CONTENU

L'analyse de l'emploi a permis de circonscrire et de comprendre l'ensemble des aspects de l'emploi, et d'en établir une description. C'est le début du processus devant permettre d'élaborer un instrument de mesure qui soit le plus représentatif possible de l'emploi. Idéalement, l'instrument de mesure devrait refléter parfaitement ce qui se passe dans l'emploi. Malheureusement, ce niveau de validité n'est jamais atteint en pratique, parce qu'il est tout simplement impossible de recréer exactement la réalité de l'emploi ; la seule chose qui soit exactement identique à l'emploi, c'est l'emploi lui-même. Pour cette raison, un instrument de mesure ne peut être qu'une approximation de l'emploi réel, et la qualité de cette approximation est la résultante d'un long travail de collecte d'informations, d'analyse et de jugement. Ce processus est illustré par le schéma de la figure 5.2.

À partir de l'univers de l'emploi, de tous ses éléments connus et inconnus, ont été circonscrits les éléments observables, non triviaux et pertinents pour donner lieu à une description de l'emploi. Cette description n'est qu'un sous-ensemble de l'univers de l'emploi, étant donné que plusieurs éléments de l'univers ne s'y trouvent plus. D'abord, les éléments inconnus et non observables par les analystes ont d'emblée été exclus. Ensuite, les éléments jugés triviaux et non pertinents ont été éliminés. Si l'analyse de l'emploi s'est déroulée parfaitement, la description de l'emploi est au mieux un sous-ensemble représentatif de l'univers de l'emploi. S'il s'est glissé des erreurs au cours de l'analyse, cette description de l'emploi devient un sous-ensemble biaisé. Dans la figure 5.1, cette perte d'information lors du passage de l'univers vers la description de l'emploi est illustrée par le carré de cette dernière qui est plus petit que celui de l'univers.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 5.2
PROCESSUS DE DÉFINITION DU CONTENU DE L'INSTRUMENT DE MESURE À PARTIR DE L'EMPLOI



Cependant, la texture est restée la même, étant donné que la description de l'emploi est censée être de même nature que l'univers, c'est-à-dire constituée des mêmes aspects observables. Les autres éléments de cette figure seront traités plus loin, à mesure que les étapes suivantes seront franchies.

ÉTAPE 3. SPÉCIFICATION DU DOMAINE OU DU SOUS-DOMAINE À MESURER

La présente étape consiste à définir le domaine de contenu à mesurer par l'instrument (voir tableau 5.1, étape 3). En sélection du personnel, le domaine de contenu correspond aux qualités recherchées chez le candidat ; l'ensemble de ces qualités porte parfois le nom de profil d'exigences, profil du candidat idéal ou encore d'énoncé de qualités. Définir le domaine ou les sous-domaines à mesurer implique deux

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

actions : choisir les éléments de l'emploi à retenir, puis les transposer de manière à fournir une définition claire du domaine ou du sous-domaine en question.

CHOIX

Éléments critiques. Il faut d'abord choisir quels éléments ou tâches de l'emploi (selon le niveau d'analyse retenu précédemment pour l'aspect « ce qui doit être effectué ») feront partie du domaine ou du sous-domaine de contenu à mesurer par l'instrument[6]. Les spécialistes et les autorités affirment que *les éléments critiques de l'emploi doivent être choisis pour être inclus dans l'instrument de mesure et, réciproquement, seulement ces éléments critiques peuvent être choisis* (SIOP, 1987, p. 23 ; EEOC, 1978, section 1607.14-C-2). Malheureusement, ces mêmes autorités n'ont pas précisé le sens du terme *critique*. Flanagan (1954) a grandement contribué à populariser ce terme par sa célèbre technique d'analyse de l'emploi, appelée la méthode des incidents critiques. On y apprend que l'expression critique est utilisée pour qualifier un élément qui a un effet important sur la réalisation des objectifs de l'emploi ; cet effet peut être direct ou indirect, à court, moyen ou long terme. L'effet doit être suffisamment important pour avoir un impact considérable sur la réalisation des objectifs de l'emploi. Signalons que l'impact peut être positif ou négatif, pourvu que son importance soit significative.

Selon cette définition, la détermination de ce qui est critique est forcément subjective et relève éventuellement de l'application d'un jugement ou d'une échelle de valeurs. Certains spécialistes préfèrent définir le caractère critique d'un élément de l'emploi en recourant à des critères plus objectifs comme la fréquence, la proportion de temps consacré, le degré de difficulté, la conséquence d'une erreur ou la formation nécessaire pour l'accomplir (Gavin, 1977 ; Schmitt et Chan, 1998). Malgré cela, une dose de jugement est toujours requise lors de la sélection des éléments qui constitueront le contenu à mesurer (Nunnally et Bernstein, 1994), d'où la nécessité de recourir à plusieurs experts de l'emploi (SME) pour juger de l'importance des divers éléments de l'emploi (Gatewood et Feild, 1998).

6. Dans la plupart des ouvrages, l'analyse de l'emploi est centrée sur « ce qui doit être effectué ». Nous verrons plus loin que les autres aspects de l'emploi ont aussi leur importance. Les aspects « ce qui doit en résulter » et « équipements et technologies » pourraient aussi être soumis à ce choix.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Cibler un sous-domaine. Il peut arriver que l'on désire se limiter à un seul sous-domaine de l'emploi. Pour un emploi de professeur d'université par exemple, les tâches se subdivisent en quatre champs de responsabilités : enseignement, recherche, administration universitaire et services à la collectivité. Lors de l'élaboration du processus d'embauche, il pourrait être décidé de concevoir un instrument de sélection, en l'occurrence une simulation, ne portant que sur le sous-domaine de l'enseignement et les capacités pédagogiques. Pour des postes de gestion, un test comme l'épreuve du courrier pourrait être soumis à des candidats pour évaluer seulement leur façon de résoudre des problèmes. Dans ce cas, le sous-domaine de contenu mesuré serait celui de la résolution des problèmes de gestion.

TRANSPOSITION

Dans un contexte de sélection et de promotion du personnel, il a été établi que, dans une stratégie de validation basée sur le contenu, le domaine de contenu doit être défini par rapport 1) aux comportements dans l'emploi, 2) aux résultats produits par ces comportements ou 3) aux connaissances ou aux habiletés nécessaires à ces comportements (voir chapitre 2, section « Détermination du domaine de contenu en contexte de gestion des ressources humaines »). Voyons comment transposer les renseignements recueillis et consignés dans la description d'emploi, puis choisis pour former le domaine ou le sous-domaine de contenu. Si l'on désire traduire le domaine de contenu en termes de comportements, il est clair que ces renseignements seront regroupés principalement sous l'aspect « ce qui doit être effectué » (voir tableau 5.2). Les autres aspects pourront fournir des compléments d'informations utiles. Si l'on préfère une transposition en termes de **résultats** produits par ces comportements, il faudra préciser avant tout les aspects « ce qui doit en résulter », « normes formelles de rendement » et « ce qui doit être effectué ». Si ce sont les **caractéristiques individuelles sous-jacentes** qui sont visées, il faudra alors les inférer à partir de toutes les informations accumulées dans la description de l'emploi. Bien entendu, la rubrique « exigences requises » pourra être d'une certaine utilité, mais sans plus. Notons que les renseignements relatifs à l'aspect « contexte de l'emploi » seront très utiles lorsque viendra le temps d'élaborer les conditions d'application de l'instrument de mesure.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Transposition et niveau d'inférence. La transposition des éléments choisis doit se faire en réduisant au minimum le recours à l'inférence, si le processus d'élaboration de l'instrument de mesure est celui de la validation basée sur le contenu ; c'est l'une des conditions d'application les plus importantes. En effet, une telle stratégie de validation repose essentiellement sur le jugement d'experts qui doivent évaluer jusqu'à quel point l'instrument reflète le domaine à mesurer, à savoir les comportements dans l'emploi, les résultats produits ou les caractéristiques individuelles requises. Or, plus l'instrument est identique à l'emploi et mesure des aspects observables, plus la démonstration des experts est directe et facile à établir. Par conséquent, la façon de réduire le niveau d'inférence consiste à demeurer le plus près possible de la description de l'emploi, basée elle-même sur des éléments observables et vérifiables (voir chapitre 2, section « Conditions d'application de la validité basée sur le contenu »).

A) Transposition, sans inférence, en comportements ou en résultats

La première façon d'effectuer la transposition des éléments choisis de l'emploi en un domaine de contenu à mesurer est de spécifier ce dernier en termes de comportements (gestes, tâches ou responsabilités) ou en termes de résultats ou de produits de ces comportements. De cette manière, la transposition est **directe** et **sans inférence**. *Les aspects observables de l'emploi sont retenus tels quels, sans qu'il y ait de véritable transposition*. Pour un emploi de secrétaire, par exemple s'il faut utiliser un certain logiciel pour la rédaction de textes en français et en anglais, le domaine à mesurer est défini par la rédaction de textes en français et en anglais à l'aide de ce logiciel. Dans un autre exemple rapporté par Guion (1977), le domaine de contenu d'un examen de compréhension de texte pour un emploi donné est constitué de toutes les tâches des deux premières semaines de travail impliquant des textes et de toutes les questions pouvant être posées sur ces textes.

Voyons un autre exemple. Le conseil municipal désire recourir à une entrevue structurée pour procéder à la sélection du futur directeur général de la Ville. L'analyse de l'emploi a déjà permis de regrouper les principales tâches d'un directeur en six grandes responsabilités : 1) la gestion des budgets, 2) la gestion de la réglementation, 3) la gestion des plaintes et des réclamations, 4) la gestion des programmes et des projets, 5) la gestion interne et la supervision du personnel et

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

6) la gestion des communications (revoir la figure 2.1). Pour le conseil, ces responsabilités constituent le domaine à mesurer. L'entrevue sera composée de quatre à cinq mises en situation par responsabilités (appelées dimensions du domaine) et permettra d'évaluer les compétences des candidats pour chacune de ces dimensions.

Dimensions génériques du rendement. Procéder directement, sans transposition, demande tout de même un certain travail d'analyse, ne serait-ce que pour regrouper les comportements ou les résultats observés dans l'emploi dans des dimensions plus larges, comme les responsabilités ou les fonctions. Les travaux de Campbell peuvent être fort utiles pour guider ces regroupements (Campbell, 1990 ; Campbell *et al.*, 1993). Cet auteur propose une structure des dimensions de la performance au travail composée de huit dimensions génériques. Un tel modèle peut se révéler une aide précieuse pour quiconque veut mesurer le rendement (voir tableau 5.4)

Ensemble, ces huit dimensions sont suffisantes pour définir la performance dans tous les postes existants. Cependant, tous les facteurs ne s'appliquent pas dans tous les emplois et peuvent exiger des ajustements. Par exemple, tous les emplois ne comportent pas de supervision ou des tâches administratives. Dans le cas des militaires non gradés de l'armée américaine, Campbell, McHenry et Wise (1990) retiennent cinq dimensions pour définir le rendement, dont quatre sont semblables aux facteurs généraux du modèle : 1) efficacité dans les tâches techniques spécifiques à leur poste (p. ex., un opérateur de char d'assaut doit pouvoir conduire son véhicule, armer le canon, tirer avec précision, etc.), 2) efficacité dans les tâches générales (premiers soins, maniement des armes de base, etc.), 3) effort et persévérance dans l'adversité, 4) discipline personnelle et 5) bonne condition physique. Enfin, Campbell (1990) croit que les trois dimensions « tâches techniques et spécifiques », « effort » et « discipline personnelle » sont des facteurs importants de rendement pour tous les emplois[7].

7. Hunt (1996) a également proposé un modèle générique du rendement, qui ne s'applique que pour les employés non spécialisés. En voici les principaux facteurs : 1) laborieux, 2) méticuleux, 3) flexible face aux horaires, 4) assidu, 5) respect des règlements, 6) rebelle et insoumis, 7) vol et 8) usage d'alcool et de drogue. Plusieurs des dimensions de Hunt, comparativement à celles de Campbell, visent davantage des qualités personnelles, donc elles sont moins utiles à ce stade-ci de notre démarche.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 5.4
DIMENSIONS GÉNÉRIQUES DU RENDEMENT

Dimension	Exemples
Efficacité dans les tâches techniques spécifiques et au cœur du poste (*job-specific task proficiency*).	Fabriquer des armoires. Faire du traitement de texte. Concevoir un logiciel. Conduire un autobus dans la circulation urbaine. Développer un système de sélection du personnel.
Efficacité dans les tâches générales par toute personne dans une organisation ou au moins appartenant à une famille d'emplois (*non-specific task proficiency*).	Les professeurs d'université doivent, en plus de maîtriser leur champ disciplinaire, savoir enseigner, analyser une demande d'admission, conseiller les étudiants, etc. Les militaires doivent maîtriser les premiers soins ou les rudiments de l'orientation, etc.
Efficacité dans les communications écrites et orales (*written and oral communication task*).	Rédiger une lettre, un rapport. Décrire une situation verbalement. Faire une présentation orale.
Effort soutenu et approprié en toute circonstance (*demonstrating effort*).	Démontrer des efforts constants, chaque jour. Persévérer, même lorsque les conditions sont difficiles. Fournir un effort additionnel lorsque requis.
Discipline personnelle (*maintaining personal discipline*).	Éviter l'alcool et les drogues. Respecter les règlements. Ne pas s'absenter de manière abusive.
Appui et aide au travail en équipe et à celui des pairs (*facilitating peer and team performance*).	Soutenir et encourager ses collègues. Aider les collègues à résoudre un problème au travail. Agir comme mentor, servir d'exemple à ses collègues.
Supervision et leadership auprès des subordonnés au moyen de comportements d'influence et d'interactions interpersonnelles (*supervision/leadership*).	Fixer des objectifs à ses subordonnés. Montrer des méthodes efficaces à ses subordonnés. Encourager les efforts de ses subordonnés. Récompenser ou punir ses subordonnés.
Tâches administratives distinctes de la supervision (*management/administration*).	Organiser les personnes et les ressources. Suivre l'avancement d'un dossier ou d'un projet. Contrôler les dépenses. Obtenir des ressources additionnelles.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

B) Transposition, avec inférence, en caractéristiques individuelles sous-jacentes

La deuxième façon, plus indirecte, consiste à spécifier le domaine à mesurer en inférant les exigences requises par l'emploi. On définit alors le domaine de contenu *en identifiant les connaissances, les habiletés et les autres caractéristiques que doit avoir une personne pour accomplir les tâches reliées à un emploi donné*. Comme le domaine ou le sous-domaine de contenu à mesurer ne sont pas définis dans les mêmes termes que la description de l'emploi, il y a forcément un certain **degré d'inférence** pour passer des éléments de l'emploi (des comportements, des résultats, des normes formelles de rendement, etc.) à des caractéristiques individuelles requises (des connaissances, des habiletés, des aptitudes, etc.). Le degré d'inférence peut être faible ou élevé selon le type de caractéristiques individuelles relevées.

KSAO. En analyse d'emploi, les caractéristiques individuelles peuvent être regroupées en fonction de quatre types de caractéristiques, soit ce qu'il est convenu d'appeler en anglais les KSAO (voir tableau 5.5). On y retrouve les **connaissances reliées à l'emploi** (*job knowledge* ou K), les **habiletés** (*skills* ou S), les **aptitudes** (*abilities* ou *aptitudes* ou A), et les **autres caractéristiques** (*others characteristics* ou O)[8].

La distinction entre les habiletés et les aptitudes n'est pas toujours évidente. Même si, théoriquement, les habiletés sont plus apprises et plus spécifiques que les aptitudes, il n'est pas facile de les démarquer. Par exemple, est-ce que « composer un discours » ou « communiquer de façon claire » est une aptitude ou une habileté ? Au fond, ce n'est pas si important. En pratique, il n'est pas essentiel qu'une caractéristique soit classée dans la bonne catégorie, ce qui compte, c'est la caractéristique elle-même et le fait qu'elle soit retenue

8. La définition de ces termes peut varier selon les auteurs. Par exemple, le terme « habiletés » (*skills*) est utilisé généralement pour désigner les compétences apprises, sans préciser s'il s'agit d'habiletés cognitives ou psychomotrices. Parfois, cependant, les habiletés ne concernent que le domaine psychomoteur. Pour éviter la confusion, le tableau 5.4 précise que les compétences appartenant aux deux domaines sont incluses dans la catégorie « habiletés ». En ce qui concerne les « aptitudes », certains auteurs (Harvey, 1991) les désignent par l'expression « *abilities* ».

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 5.5
TYPES DE CARACTÉRISTIQUES INDIVIDUELLES

	Type de caractéristiques	Définition	Exemples
K	Connaissances (*job knowledges*)	Ensemble d'informations portant sur des faits, des règles ou des procédures et s'appliquant directement à la réalisation d'une activité, d'une tâche ou d'une fonction. Les connaissances sont à la base des habiletés cognitives.	Connaissance de l'orthographe, de la convention collective, d'un logiciel, des préférences de ses employés, du plan stratégique de l'organisation, etc.
S	Habiletés psychomotrices (*skills*)	Performance (compétence) observable dans la réalisation apprise d'une activité ou d'une tâche de nature physique ou motrice (c.-à-d. impliquant des mouvements du corps et des membres, l'usage de la vision, etc.).	Appliquer les freins, percer des trous dans du métal, taper sur un clavier d'ordinateur, insérer des lettres dans des enveloppes, etc.
	Habiletés cognitives (*skills* ou *abilities* selon les auteurs)	Performance (compétence) observable dans la réalisation apprise d'une activité, d'une tâche ou d'une fonction de nature cognitive.	Élaborer le budget de son service, utiliser un ordinateur pour préparer des états financiers, questionner un client pour identifier ses besoins, développer un système de classement des plaintes, etc.
A	Aptitudes cognitives et psychomotrices (*abilities* ou *aptitudes* selon les auteurs)	Capacités potentielles qui influencent l'apprentissage et l'exécution d'une activité, d'une tâche ou d'une fonction.	Aptitude verbale, capacité d'apprentissage, perception spatiale, coordination visuo-motrice, etc.
		Capacités générales que la personne possède avant l'apprentissage d'une tâche spécifique.	Capacité à comprendre des directives écrites, à faire un exposé oral, à faire des calculs de base, à décrire une situation, à écrire des phrases, etc.
O	Autres caractéristiques (*other characteristics*)	Traits de personnalité, besoins, valeurs, etc.	Introversion, stabilité émotionnelle, besoin d'estime, honnêteté, etc.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

pour constituer le domaine à mesurer (Gatewood et Feild, 1998). Il est également intéressant de rappeler que des connaissances sont toujours impliquées dans les habiletés cognitives (réaliser un tableau à l'ordinateur, faire un rapport sur le bilan des opérations, remplir un bordereau pour commander du matériel, etc.) et psychomotrices (conduire une automobile, souder à l'arc électrique, assembler des pièces, etc.)[9].

Modèles ou typologies portant sur les caractéristiques individuelles. Lorsque vient le temps de relever les diverses caractéristiques individuelles requises pour un emploi, il est plus efficace de se servir de théories éprouvées que de tenter de réinventer la roue. Pour les **aptitudes** (A), Fleishman et Reilly (1992) fournissent sans doute la typologie la plus complète qui soit sur le monde du travail. Elle comprend 52 aptitudes génériques, groupées en quatre catégories : cognitive, psychomotrice, physique et sensorielle/perceptuelle. Quant aux **traits de personnalité** (O), le modèle des cinq facteurs de Digman (1990) émerge de plus en plus en psychologie industrielle. Ces cinq facteurs de la personnalité peuvent être définis par les qualités suivantes : 1) extraverti, sûr de soi, démonstratif, 2) agréable, aimable, empathique, 3) consciencieux, fiable, méticuleux, 4) émotionnellement stable, ou maître de ses émotions et ses humeurs et 5) curieux, imaginatif, aimant jongler avec les idées. Pour ce qui est des **intérêts professionnels** (O), le modèle RIASEC de Holland (1973) fournit six grandes dimensions : réaliste, investigateur, artistique, social, entrepreneur et conventionnel. Campbell (1990) mentionne d'autres taxonomies intéressantes, notamment celle de Peterson et Bownas (1982) pour les aptitudes, la personnalité et les intérêts professionnels, de Snow (1989) pour les aptitudes et de Owens et Schoenfeldt (1979) pour les données biographiques. Il n'y a pas vraiment de typologies reconnues pour les connaissances (K) et les habiletés (S), chaque emploi étant particulier. Par contre, on retrouve souvent les mêmes habiletés dans les profils d'exigences élaborés par les grandes organisations ou par les cabinets-conseils en sélection du personnel (Pettersen, 1991 ; Spencer et Spencer, 1993).

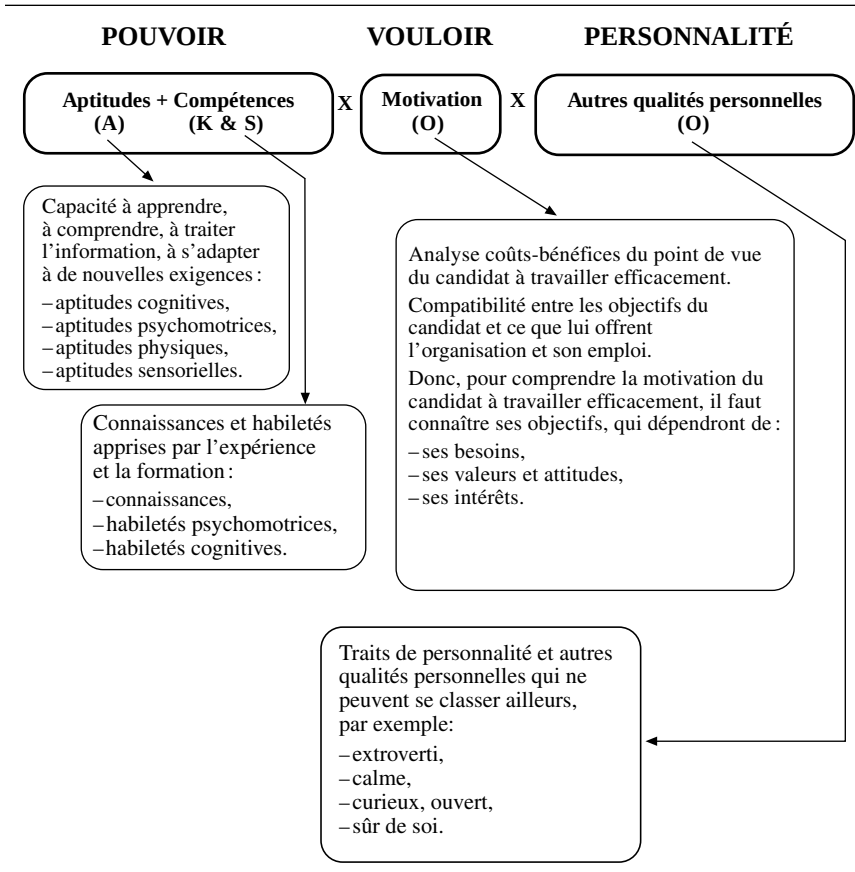
9. Pour une présentation plus élaborée des caractéristiques individuelles et de leur définition, le lecteur est invité à consulter Dunnette (1976), Harvey (1991), Guion (1998), Pettersen et Jacob (1992) ou Slivinsky et Miles (1996). D'autres exemples de typologie ou de grilles conçues pour la sélection du personnel cadre ou professionnel sont proposés par les auteurs suivants : Baehr (1993), Boyatzis (1982), Pettersen (1991), Spencer et Spencer (1993).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Un schéma d'ensemble des composantes du comportement au travail. Dans un contexte de gestion des ressources humaines, l'ensemble des caractéristiques du modèle KSAO peut être articulé de manière à rendre compte du comportement d'un individu au travail. On retrouve à la figure 5.3 un exemple de schéma utile à cet égard (adapté de Pettersen et Jacob, 1992). On remarque d'abord que les KSAO servent à définir trois grandes composantes qui influent sur le comportement au travail : le pouvoir, le vouloir et la personnalité. Le **pouvoir**, ou ce qu'une personne est capable de faire, dépend de ses aptitudes (A) et de ses compétences (KS), ces dernières étant composées des connaissances et des habiletés que la personne a acquises

Figure 5.3
COMPOSANTES DU COMPORTEMENT AU TRAVAIL



© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

grâce à sa formation et à son expérience. Vient ensuite le **vouloir**, qui concerne cette fois ce que la personne veut faire ou ce qu'il est convenu d'appeler la motivation. Les besoins, les valeurs, les attitudes et les intérêts sont les variables à la base des motifs et des buts qui poussent l'individu à agir ; ces motifs et ces buts constituent à leur tour la source de la motivation. Finalement, il y a les traits de **personnalité** et les autres qualités personnelles qui teintent les façons d'être ou d'agir de l'individu. Les variables de la motivation et de la personnalité sont classées indistinctement dans la catégorie « O » du modèle KSAO.

Définition des dimensions du domaine lorsqu'il s'agit des KSAO. La définition du domaine de contenu à mesurer doit être *suffisamment claire pour permettre à des experts compétents de statuer sur la relation entre les items ou les questions d'un instrument et le domaine qu'ils sont censés représenter* (*Standards*, 1999, article 3.2). Lorsque le domaine de contenu est défini par des comportements ou des résultats provenant directement de l'emploi (c.-à-d. sans inférence), la définition du domaine est sans équivoque par rapport à l'emploi et la relation entre l'instrument et l'emploi peut être vérifiée aisément. En revanche, lorsque le domaine de contenu est composé de caractéristiques sous-jacentes (KSAO) qui ont été inférées, il est plus difficile de statuer sur la relation entre l'instrument et l'emploi ; une définition claire, sans ambiguïté du domaine de contenu est alors essentielle.

Le terme **dimension** est souvent utilisé pour désigner chaque caractéristique, habileté, aptitude ou ensemble de connaissances retenues pour composer le domaine à mesurer. La définition du domaine de contenu est obtenue par la définition de chacune de ces dimensions. Bien que diverses règles aient été proposées (Gatewood et Feild, 1994), la clarté et la spécificité sont les plus importantes. Voici quelques exemples de dimensions et leur définition :

– Connaissance en entrevue de sélection, incluant la préparation, la formulation des questions, les règles de conduite durant l'entrevue, la prise de notes et l'interprétation des réponses.

– Connaissance des quatre opérations de base en arithmétique, incluant l'usage des décimales.

– Habileté à taper du texte en français, au rythme de 50 mots à la minute sans erreur.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

– Habileté à faire des présentations orales structurées et qui captent l'attention devant des groupes de 20 à 50 personnes.
– Aptitude à comprendre et à suivre des directives écrites ou orales.
– Aptitude à comprendre les autres et à montrer de l'empathie à leur égard dans des situations de crise.

Un exemple de définition de cinq dimensions de gestion chez des cadres mesurées à l'aide d'une épreuve du courrier est présenté au tableau 5.6[10].

À cette étape, le processus vise à fournir une définition **conceptuelle** du domaine ou du sous-domaine à mesurer : une définition à l'aide de mots et de concepts (Kerlinger, 1986). Ce n'est qu'à des étapes ultérieures qu'une définition plus **opérationnelle** du domaine sera élaborée, c'est-à-dire une définition qui précise de quelle manière ce domaine sera mesuré. Cette définition sera en fait l'instrument de mesure et ses diverses composantes[11].

Importance relative des dimensions. Il est recommandé d'évaluer l'importance relative de chacune des dimensions en fonction de l'emploi. La plupart des auteurs recommandent de recourir à un échantillon d'experts qui verra alors à coter les dimensions au moyen d'échelles et de méthodologies plus ou moins complexes (Gatewood et Feild, 1998 ; Schmitt et Chan, 1998).

---

10. Commission de la Fonction publique du Canada, *Renseignements sur l'Exercice «* In-Basket *» pour la gestion (810)*, Ottawa, Centre de psychologie du personnel.
11. Cela rejoint la position de Cronbach (1990) lorsqu'il affirme que toute définition du domaine de contenu doit absolument couvrir les trois aspects suivants : 1) les tâches, les stimuli ou les situations, 2) les directives ou les ordres adressés au sujet et 3) les réponses que l'observateur ou le correcteur devra compter. Cette correspondance entre les trois aspects relevés par Cronbach concernant le domaine de contenu et les composantes d'un instrument de mesure est normale. En effet, si les stimuli et les conditions sont des aspects qui font partie de la spécification même de la nature de la performance ou des caractéristiques individuelles sous-jacentes à mesurer, ces mêmes spécifications doivent se retrouver dans l'instrument servant à recueillir un échantillon de cette performance ou de ces caractéristiques individuelles sous-jacentes. Quant à la performance recherchée ou à la manifestation des caractéristiques individuelles sous-jacentes, il est tout aussi naturel que ces éléments se retrouvent consignés dans le processus d'observation et de correction de l'instrument de mesure, compte tenu que ce processus a précisément pour fonction d'encadrer le correcteur dans sa reconnaissance des éléments de réponse attendus.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 5.6
DÉFINITION DES CINQ DIMENSIONS QUI COMPOSENT LE DOMAINE DE CONTENU MESURÉ PAR UNE ÉPREUVE DU COURRIER (*IN-BASKET TEST*)

- **Planifier**

  Capacité de planifier et de mettre en œuvre des solutions ou des interventions efficaces pour résoudre des problèmes organisationnels touchant les programmes, les projets, les finances ainsi que les exigences du personnel et du public. Cette capacité regroupe diverses activités, dont la planification de réunion pour discuter de problèmes et de plans d'action ; l'amorce de recherche ; la rédaction d'ordres du jour, de rapports, de discours et d'autres documents ; la recommandation de solutions de rechange ainsi que l'amélioration de programmes et de systèmes.

- **Diriger**

  Capacité de diriger et de conseiller efficacement le personnel ainsi que de mettre en place des mécanismes de rétroaction nécessaires pour suivre de près les activités et les coûts.

- **Analyser**

  Capacité d'analyser et d'évaluer des problèmes touchant le personnel, les opérations, l'organisation et la gestion du budget. Cette capacité se manifeste par la détermination des causes sous-jacentes aux problèmes, d'éléments communs, d'interrelations, de solutions aux problèmes, de questions plus vastes et de répercussions sur l'organisme.

- **Responsabiliser**

  Capacité de répartir efficacement le travail des employés et des employées en tenant compte des capacités et des responsabilités de chacun, de l'attribution efficace des fonctions opérationnelles et des méthodes de contrôle en place.

- **Organiser**

  Capacité d'organiser votre travail et vos activités par la gestion systématique de vos heures de travail. Cela implique aussi votre capacité d'établir des méthodes qui assureront à long terme l'utilisation maximale des heures de travail et des ressources humaines. Vous devez pouvoir échelonner votre travail personnel de façon systématique, tenir un calendrier de vos activités et porter un jugement critique sur votre propre emploi du temps.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Relation entre les KSAO et les éléments de l'emploi. La plupart des auteurs recommandent que la relation entre les KSAO et les éléments de l'emploi soit vérifiée après coup, notamment par un nouvel échantillon d'experts (SME). Différentes méthodes peuvent être utilisées (Gatewood et Feild, 1998 ; Guion, 1998 ; Schmitt et Chan, 1998). Une façon de maintenir la relation KSAO-emploi est proposée plus loin (voir la section « Structure du domaine à mesurer », au paragraphe « Structures intégrant des KSAO »).

Caractéristiques individuelles et degré d'inférence. Revenons au degré d'inférence nécessaire pour passer de l'emploi au domaine de contenu à mesurer. Se référant aux EEOC (1978) et aux *Standards* (1985), Harvey (1991) souligne que les connaissances et les habiletés spécifiques à l'emploi (KS) sont très près des comportements observables dans l'emploi, alors que les aptitudes et les autres caractéristiques (AO) sont reliées plus indirectement, sinon pas du tout, aux comportements de l'emploi. Comme elles sont des construits, les AO ne sont pas directement observables, mais plutôt inférées à partir de styles ou de modèles (*patterns*) observés dans les comportements.

Mais même si les aptitudes et les autres caractéristiques (AO) sont traduites éventuellement en comportements observables, tout comme les connaissances et les habiletés (KS), il ne s'agit pas des mêmes agrégats de comportements. Ainsi, les **connaissances reliées à l'emploi et les habiletés cognitives (K)** sont spécifiées dans les mêmes termes que dans l'emploi (connaissance de l'orthographe, maîtrise de tel logiciel, rédaction d'une lettre, etc.). Les **habiletés (S)** sont aussi définies par des comportements observables qui sont souvent identiques à ceux décrits dans l'emploi (faire fonctionner tel appareil, opérer une scie radiale, conduire une réunion, etc.). Il est exact que les habiletés et les connaissances ne sont pas en soi observables directement ; cependant leur relation avec les éléments de l'emploi est très étroite. Par conséquent, le degré d'inférence nécessaire pour transposer les éléments de l'emploi aux connaissances et aux habiletés du domaine de contenu est minime.

Il en est tout autrement des **aptitudes (A) et des autres caractéristiques (O)**. Ces facteurs englobent des modèles de comportements plus généraux et plus éloignés des éléments de l'emploi. Pour les aptitudes, par exemple, répondre à des items de tests psychométriques portant sur la perception en trois dimensions, sur des synonymes et

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

des antonymes ou sur la recherche de figures identiques ne représente pas tout à fait le genre de comportements que l'on retrouve dans le monde du travail. Il en va de même pour les inventaires de personnalité, de besoins ou de valeurs, où la personne doit choisir parmi un ensemble d'énoncés ceux qui la caractérisent le mieux. Répondre à un test ou un inventaire et agir réellement dans une situation de travail sont deux choses fort différentes[12].

Caractéristiques individuelles et application de la validation basée sur le contenu. Il est évident, en gestion des ressources humaines, que la validation de contenu s'applique moins aux instruments qui mesurent les aptitudes et les autres caractéristiques personnelles (AO). Si le domaine à mesurer est spécifié par les **connaissances** et les **habiletés** propres à l'emploi (KS), ou encore, par les comportements observés dans l'emploi (éléments, tâches et responsabilités) ou les résultats ou les produits de ces comportements, la validation de contenu pourra aisément être appliquée (Cronbach, 1970 ; EEOC, 1978, 1607.14.C.1 ; Schneider et Schmitt, 1986) et être suffisante pour établir la validité de l'instrument (SIOP, 1987). En revanche, si le domaine (ou sous-domaine) à mesurer est défini en fonction d'**aptitudes** ou d'**autres caractéristiques (AO)** plus abstraites, la validité basée sur le contenu de l'instrument sera difficile à établir (Arvey et Faley, 1988).

La Society for Industrial and Organizational Psychology (1987) et le United States Civil Service Commission (Gavin, 1977) reconnaissent que des spécialistes appliquent la validation basée sur le contenu à des instruments mesurant de tels éléments plus généraux ou plus abstraits. Toutefois, la méthode de validation basée sur le contenu est rarement appropriée et ne suffit pas à démontrer la validité d'un tel instrument

12. Signalons que les connaissances et les habiletés (KS) se distinguent aussi des aptitudes et des autres caractéristiques (AO) en ce qui concerne leur stabilité. Les connaissances et les habiletés (KS) sont apprises, c'est pourquoi elles peuvent être améliorées par l'expérience ou par la formation. Les aptitudes (A) et les autres caractéristiques (O), comme les traits de personnalité, les besoins fondamentaux et les valeurs, sont considérées comme étant relativement stables chez un adulte, du moins beaucoup plus stables que les connaissances et les habiletés. Conséquemment, elles (AO) sont plus difficiles à modifier.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

(*Standards*, 1999, article 14.9 ; EEOC, 1978, N° 1607.14.C.1)[13]. Dans ce cas, il faut recourir à d'autres stratégies de validation (Gatewood et Feild, 1998 ; Gavin, 1977 ; SIOP, 1987).

Évidemment, il n'est pas facile d'établir une nette distinction entre une habileté et une aptitude, entre une caractéristique spécifique à l'emploi et une caractéristique plus générale. Cependant, plus le contenu de l'instrument est défini en des termes éloignés de l'emploi, plus il est souhaitable de faire appel à des experts qualifiés (SME) et nombreux pour identifier les KSAO, afin de restreindre la probabilité des erreurs de jugement inhérentes à l'accroissement du degré d'inférence. De plus, on devra déterminer les KSAO sous-jacentes seulement après avoir circonscrit les éléments critiques de l'emploi (Thompson et Thompson, 1982).

LIMITES À ÊTRE TRÈS SPÉCIFIQUE À L'EMPLOI

Demeurer le plus près possible des éléments observables de la description de l'emploi afin de réduire au minimum le degré d'inférence comporte certains pièges.

Vulnérable aux changements et non transférable à d'autres emplois. Plus le contenu à mesurer reflète en tout point un emploi, moins l'instrument peut être généralisé (SIOP, 1987, p. 20). Il sera plus difficile pour cet instrument très spécifique de conserver sa validité à la suite du moindre changement qui surviendra dans l'emploi. Il sera également plus difficile d'étendre la validité de contenu de cet instrument à d'autres emplois d'une même famille. La généralité et la spécificité d'un instrument de mesure sont les deux extrémités d'un même continuum. C'est à l'organisation ou au concepteur de choisir le degré de généralité ou de spécificité que doit avoir l'instrument. Par exemple, si plusieurs emplois sont visés, il faut explorer la possibilité de définir le contenu à mesurer à l'aide d'éléments assez généraux, comme les connaissances et les habiletés (KS), afin d'englober

13. Pour cause d'inférence également, cette stratégie de validation ne convient pas non plus aux instruments dont les stimuli ou les conditions d'administration ne ressemblent pas à ceux de l'emploi (Gatewood et Feild, 1998). Bien que pour des raisons différentes, la validité de contenu ne constitue pas non plus une méthode appropriée pour démontrer la validité d'un instrument qui mesure des connaissances ou des habiletés qui seront acquises plus tard par le candidat une fois en emploi (Equal Employment Opportunity Commission *et al.*, 1978, 1607.14.C.1).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

ce qui est commun à l'ensemble de ces emplois, sans nécessairement aller jusqu'au niveau des aptitudes et des autres caractéristiques personnelles (AO). Lors de la définition du domaine ou des sous-domaines à mesurer, il est essentiel que le degré de généralité recherché dans l'instrument de mesure ait déjà été établi (SIOP, 1987, p. 20). Dans la présente démarche, cela a été fait à l'étape 1 sur les finalités de l'instrument de mesure.

Ne permet pas de comprendre les facteurs de réussite dans l'emploi. Il est difficile d'expliquer la réalisation des tâches et, par conséquent, d'identifier les facteurs de succès et d'échec, lorsque la transposition ne recourt pas aux caractéristiques sous-jacentes (KSAO). Ce problème peut survenir lorsque la transposition est faite en termes de comportements ou de résultats. Par exemple, les dimensions « capacité à travailler sous tension » ou « capacité à faire face à des clients récalcitrants » n'indiquent pas ce qui est requis pour réussir les tâches qui composent ces dimensions.

Source possible de redondance entre les dimensions du domaine de contenu. Le fait de ne pas identifier les caractéristiques sous-jacentes (KSAO) peut amener un autre problème (Gatewood et Feild, 1998), celui de la redondance entre les dimensions du domaine de contenu. Prenons l'exemple de la dimension « capacité à faire face à des clients récalcitrants ». Il est probable que des KSAO sous-jacents (p. ex., connaissances des techniques de négociation ou habiletés à trouver des solutions gagnant-gagnant) le soient aussi pour une autre dimension comme « capacité à mobiliser les membres de son équipe ». À la limite, si chaque tâche de l'emploi devient une dimension à évaluer, c'est considérer que chacune d'elles est différente du point de vue des caractéristiques sous-jacentes, ce qui est rarement le cas.

Risque de préjudice. Il n'est pas raisonnable qu'un instrument de mesure soit spécifique à un point tel qu'il soit réussi seulement par le candidat ayant déjà occupé l'emploi ou ayant reçu une formation propre à cet emploi (Arvey et Faley, 1988). Par exemple, il semblerait discutable a priori qu'un examen d'informatique utilise un logiciel ayant été modifié localement ou porte sur des connaissances précises concernant les rouages administratifs d'un service. Il est recommandé de déterminer quelles sont les connaissances et les habiletés

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

qu'une personne doit posséder avant d'entrer en fonction et de restreindre le domaine de contenu en conséquence. Nous abordons cet aspect à la section suivante.

FRONTIÈRES DU DOMAINE À MESURER

Un domaine implique des frontières (Murphy et Davidshofer, 1988), ce qui permet de déterminer si un élément est à l'intérieur ou à l'extérieur du domaine. En définissant le contenu du domaine ou des dimensions à mesurer, les frontières ont été jusqu'à un certain point tracées. Par exemple, à partir de la définition des cinq dimensions composant le domaine de contenu à mesurer par cette épreuve du courrier (voir tableau 5.6), on a une idée du contenu de ces capacités de gestion, délimitant du coup leurs frontières.

Cependant, les frontières seront mieux définies s'il est indiqué que certains éléments ne devraient pas se retrouver à l'intérieur du domaine (SIOP, 1987, p. 20). Par exemple, supposons que l'épreuve du courrier contient une trentaine de problèmes. Supposons également que parmi les personnes qui postulent, il y a des candidats de l'interne qui travaillent actuellement au gouvernement et des candidats de l'externe qui proviennent d'autres organisations. Si les problèmes étaient des cas réels déjà survenus et dont le dénouement (la solution) est connu des candidats à l'emploi du gouvernement, ces derniers seraient fortement avantagés. Leurs réponses dépendraient de leurs connaissances de ces événements et donneraient lieu à une évaluation biaisée de leurs véritables habiletés en gestion. Pour cette raison, il est plus sage d'exclure du domaine à mesurer toute référence à des événements survenus au gouvernement et de s'en tenir plutôt à des problèmes plus généraux ou pour lesquels aucun candidat ne possède d'informations privilégiées. De cette façon, les réponses des candidats refléteront davantage le domaine à mesurer et non leurs connaissances d'événements particuliers.

Il est recommandé de restreindre le domaine de contenu aux connaissances et aux habiletés qu'une personne doit posséder avant d'entrer en fonction (SIOP, 1987, p. 22 ; *Standards*, 1999, article 14.8) et d'éliminer les éléments qu'une personne peut apprendre lors d'une brève période d'adaptation à l'emploi (Schneider et Schmitt, 1986). Exiger des connaissances singulières, que seuls les initiés à l'emploi peuvent détenir, est trop spécifique pour un instrument de mesure.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Autant que possible, le domaine à mesurer devrait porter sur des caractéristiques suffisamment stables du candidat, qui ne vont pas changer substantiellement avec le temps (*Standards*, 1999, article 14.8).

La jurisprudence québécoise en matière des clauses d'ancienneté dans les conventions collectives relativement aux promotions confirme cette position ; les tests et examens ne doivent pas exiger une connaissance que seul un entraînement préalable à la fonction postulée aurait pu procurer (Vézina, 1979). En d'autres termes, il n'est pas nécessaire qu'un candidat à un poste soit familier avec les exigences du poste au point qu'il puisse s'acquitter des tâches s'y rapportant sans période d'apprentissage.

STRUCTURE DU DOMAINE À MESURER

La validité d'un instrument dépend du degré de similitude entre son contenu et le domaine à mesurer. Une façon d'y arriver est de subdiviser le domaine ou le sous-domaine à mesurer en ses principaux aspects, puis d'assigner à l'instrument de mesure le nombre de stimuli (items de test, questions d'entrevue, tâches dans une simulation, etc.) de manière à représenter chacune de ces subdivisions (Cronbach, 1990 ; Helmstadter, 1964). Les stimuli de l'instrument de mesure sont alors appariés (*match*) aux subdivisions du domaine visé (SIOP, 1987, p. 23). L'appariement par subdivision permet d'éviter que l'instrument comporte trop d'items pour un aspect du domaine (p. ex., planifier) et pas assez pour un autre (p. ex., diriger)[14]. On dit alors que les subdivisions confèrent une structure au domaine. L'appariement est utile aussi pour les autres composantes de l'instrument, particulièrement pour le processus d'évaluation qui contient les réponses attendues et leur pointage. En guise d'illustration, voyons quelques cas de domaines de contenu ainsi que leur structure.

Une seule dimension, mais d'innombrables manières d'apparier les stimuli aux subdivisions. Prenons l'exemple d'un directeur d'usine, chez qui l'on veut mesurer la dimension « résolution de problèmes et prise de décisions », à l'aide d'une entrevue et d'une douzaine de mises en situation. La structure de ce domaine de contenu peut être définie

14. Cette approche d'appariement par subdivision est analogue à un échantillonnage par stratification, où les sujets de l'échantillon sont appariés aux diverses strates de la population ciblée.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

par un processus comprenant les quatre opérations suivantes : 1) le diagnostic des éléments de la problématique, 2) l'identification et l'évaluation des solutions appropriées, 3) le choix de la meilleure solution et 4) la mise en application de la solution retenue. Ces quatre opérations forment une même dimension, soit les étapes du processus de « résolution de problèmes et prise de décisions ».

Comment faire en sorte que les 12 mises en situation de l'entrevue constituent un échantillon représentatif de ces 4 subdivisions ? Il y a d'innombrables manières de concevoir l'appariement des stimuli aux subdivisions. Une **première** possibilité consisterait simplement à répartir également les 12 mises en situation entre les 4 subdivisions, soit 3 problèmes par subdivision. Chacune des mises en situation ne devrait alors porter que sur la subdivision qu'elle est censée représenter. Une **deuxième** possibilité serait de retenir 12 mises en situation, chacune englobant l'ensemble des 4 opérations. Il y a **une troisième possibilité** : choisir 6 mises en situation couvrant les deux premières opérations et 6 autres pour les deux dernières opérations. Une **quatrième** possibilité pourrait mettre en jeu des mises en situation qui valent 5 points et d'autres 10 points, mais réparties de façon à ce que le total de points soit égal pour chacune des 4 subdivisions. Et ainsi de suite.

Dans cet exemple, les subdivisions (4 opérations) du domaine « résolution de problèmes et prise de décisions » sont considérées comme étant d'égale importance. On aurait pu imaginer des possibilités où l'importance des subdivisions n'est pas égale, afin de refléter les priorités de l'organisation. Bref, le nombre de possibilités pour obtenir la représentativité du domaine de contenu est théoriquement illimité.

Structures à plusieurs axes. Il se peut que les 4 opérations de la « résolution de problèmes et prise de décisions » ne suffisent pas à représenter toute la complexité du travail d'un directeur d'usine. Même si les 12 mises en situation retenues dans l'instrument couvrent adéquatement les 4 opérations qui définissent le domaine de contenu, cela ne garantit pas qu'elles soient un échantillon représentatif de toutes les situations de l'emploi de directeur d'usine nécessitant le recours à la « résolution de problèmes et à la prise de décisions ». Il peut y avoir d'autres facteurs à considérer et qui pourraient contribuer à spécifier davantage le domaine.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Par exemple, les **fonctions de l'entreprise** constituent un autre axe utile. En effet, un directeur d'usine est appelé à résoudre des problèmes et à prendre des décisions dans le cadre de ses innombrables tâches et responsabilités, lesquelles peuvent être réparties selon les diverses fonctions de l'entreprise : 1) production, 2) gestion du personnel, 3) finances et comptabilité, 4) marketing, 5) recherche et développement, 6) systèmes d'information. On pourrait ainsi créer une structure à deux axes, combinant les 4 opérations avec les 6 fonctions de l'entreprise (voir figure 5.4). Les situations-problèmes du directeur d'usine pourraient alors être réparties pour représenter simultanément les opérations et les fonctions.

Les **fonctions du management** auraient pu aussi être utilisées comme deuxième axe de spécification dans cet exemple. Les tâches et les responsabilités d'un directeur d'usine peuvent effectivement être classées selon les 4 fonctions du management : 1) la planification, 2) l'organisation, 3) la direction et 4) le contrôle (voir figure 5.5). C'est une typologie fort connue qui a fait ses preuves en gestion.

On pourrait créer une structure à trois axes, comprenant les 4 opérations, les 6 fonctions de l'entreprise et les 4 fonctions du management pour obtenir 96 subdivisions, ce qui ferait beaucoup de subdivisions à considérer pour seulement 12 mises en situation.

Figure 5.4
STRUCTURE DU DOMAINE « RÉSOLUTION DE PROBLÈMES ET PRISE DE DÉCISIONS » SELON LES OPÉRATIONS IMPLIQUÉES ET LES FONCTIONS DE L'ENTREPRISE

Fonctions de l'entreprise	Opérations			
	Diagnostic	Évaluation	Choix	Mise en application
Production				
Gestion du personnel				
Finance et comptabilité				
Marketing				
R-D				
Systèmes d'information				

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 5.5
STRUCTURE DU DOMAINE « RÉSOLUTION DE PROBLÈMES ET PRISE DE DÉCISIONS » SELON LES OPÉRATIONS IMPLIQUÉES ET LES FONCTIONS DU MANAGEMENT

Fonctions du management	Opérations			
	Diagnostic	Évaluation	Choix	Mise en application
Planification				
Organisation				
Direction				
Contrôle				

Structures intégrant des KSAO. Dans les exemples de structure précédents, le domaine de contenu « résolution de problèmes et prise de décisions » et ses divers axes de spécification visent des comportements observables dans l'emploi ou des résultats de ces comportements. Le domaine de contenu aurait pu être défini par les caractéristiques individuelles, les KSAO que doit posséder un directeur d'usine en matière de « résolution de problèmes et de prise de décisions ». Par exemple, un domaine pourrait être spécifié par les trois capacités suivantes : l'analyse, le jugement et l'esprit de décision (voir tableau 5.7).

Ces capacités constituent un exemple de dimensions éloignées de l'emploi. Malgré des définitions assez claires, relier ces capacités à ce qui est observé directement au travail exige un certain degré d'inférence. Toutefois, cela n'invalide pas les capacités retenues, mais signale simplement que la démonstration de leur pertinence par rapport à l'emploi ne pourra se faire sans recourir de nouveau à l'inférence[15]. Pour remédier à une telle situation, on aurait pu créer une structure à deux axes, l'un formé par les KSAO retenus (soit l'analyse, le jugement et l'esprit de décision), l'autre par les tâches ou leurs regroupements relevés lors de l'analyse de l'emploi (p. ex., les six fonctions de l'entreprise).

15. À moins d'emprunter une autre démarche comme la validité basée sur la relation avec d'autres variables.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 5.7
DÉFINITION DES CAPACITÉS QUI COMPOSENT LE DOMAINE « RÉSOLUTION DE PROBLÈMES ET PRISE DE DÉCISIONS »

Analyse (traitement de l'information)	Traite rapidement une grande quantité d'informations. Relève les éléments pertinents. Établit les liens entre les éléments. Distingue les faits des hypothèses, les causes des symptômes. Envisage l'ensemble des solutions possibles et leurs conséquences.
Jugement et sens pratique (contenu des solutions ou des décisions)	Tient compte de la perspective globale et non pas de quelques facettes du problème. Envisage des solutions qui tiennent compte des contraintes de l'organisation et de son environnement. Choisit la solution la plus appropriée dans les circonstances.
Esprit de décision (passage à l'action)	Cerne les priorités et établit un plan d'action. Prend des décisions et applique des solutions. S'implique dans des situations difficiles ou délicates dont les conséquences peuvent être personnellement déplaisantes.

Règles s'appliquant à une structure. Comme on peut le constater, il n'y a pas qu'une seule structure intéressante pour un domaine ou un sous-domaine en particulier. Au contraire, il y a plusieurs façons d'y arriver. Cependant, certaines règles doivent être respectées, comme avec toute typologie (Kerlinger, 1986). Premièrement, l'ensemble des axes retenus et leurs subdivisions doit permettre de rendre compte de **tout le domaine de contenu** en cause. Deuxièmement, les axes devraient idéalement être **indépendants** les uns des autres, et présenter des logiques de classification non redondantes. Par exemple, les quatre opérations de la résolution de problèmes et les six fonctions de l'entreprise constituent deux axes indépendants (voir figure 5.5). On peut trouver des problèmes servant à mesurer chacune des 24 combinaisons formées par ces 2 axes. Mais, ce n'est plus le cas si l'on remplace les fonctions de l'entreprise par les trois capacités « analyse, jugement et esprit de décision » (voir figure 5.6). Par exemple, la capacité « analyse » est proche de l'opération « diagnostic ». Toutefois, cette capacité s'applique peu à l'opération

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

« mise en application », de sorte qu'il sera difficile de trouver des problèmes mesurant la combinaison « analyse-mise en application ». La même remarque vaut pour la capacité « esprit de décision » qui s'applique mal à une opération comme « diagnostic » ou « évaluation » ; une combinaison de ces deux axes n'est pas appropriée. Leur manque d'indépendance impliquerait que, parmi les 12 cellules de la structure ainsi créée, certaines demeureraient sans items et ne pourraient pas être mesurées.

Figure 5.6
STRUCTURE DU DOMAINE « RÉSOLUTION DE PROBLÈMES ET PRISE DE DÉCISIONS » SELON LES OPÉRATIONS IMPLIQUÉES ET LES CAPACITÉS SOUS-JACENTES

Capacités	Opérations			
	Diagnostic	Évaluation	Choix	Mise en application
Analyse				?
Jugement				
Esprit de décision	?	?		

Troisièmement, les dimensions (p. ex, les trois capacités sous-jacentes) devraient être **mutuellement exclusives**, de manière à ce qu'un élément du domaine de contenu ne puisse appartenir qu'à une et une seule de ces subdivisions. Par exemple, les comportements qui définissent la capacité « analyse » ne devraient pas pouvoir se retrouver dans « jugement » ou « esprit de décision ». Cette dernière règle n'est pas toujours respectée en gestion des ressources humaines, ce qui crée de la redondance entre les évaluations des diverses dimensions chez un individu.

Table de spécifications. La délimitation des frontières a permis d'établir ce qui se trouvait à l'extérieur du domaine à mesurer. La structure permet maintenant de préciser l'intérieur du domaine, de manière à en constituer une représentation univoque, exhaustive et non biaisée. La définition du domaine à mesurer, incluant ses frontières et sa structure, servira ensuite de canevas pour élaborer l'instrument de mesure. Ce canevas est aussi appelé la **table de spécifications** ou plan du contenu (*outline of content* ; Nunnally et Bernstein, 1994).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

PROCESSUS DE DÉFINITION DU DOMAINE DE CONTENU

La troisième étape qui s'achève complète la phase 1 (voir tableau 5.1). Par rapport au processus de définition du contenu de l'instrument de mesure (voir figure 5.2), on constate que cette troisième étape a permis de passer de la description de l'emploi, effectuée à l'étape précédente, à la définition du domaine ou des sous-domaines à mesurer. Pour ce faire, il a fallu choisir parmi les éléments de l'emploi ceux qui sont critiques et pertinents au domaine à mesurer. Il est donc naturel que ce domaine de contenu soit dans presque tous les cas plus restreint que la description de l'emploi. Ce phénomène est illustré à la figure 5.2 par un contour du domaine ou du sous-domaine plus petit que celui de la description de l'emploi.

Il a fallu également transposer les éléments choisis en comportements, en résultats produits par ces comportements ou en caractéristiques individuelles sous-jacentes, tout en demeurant le plus proche possible de l'emploi et en réduisant ainsi le recours à l'inférence. Dans la figure, le changement de texture (ombre plus pâle) du domaine ou des sous-domaines de contenu à mesurer rappelle que cette transposition peut amener les éléments à ne plus nécessairement être spécifiés dans les mêmes termes que ceux de l'emploi. Ce traitement de l'information a nécessité de la part du concepteur un certain degré d'inférence ou de jugement. D'abord, il y a eu inférence lors du choix des éléments. Ensuite, si ces éléments n'ont pas été définis directement par les comportements de l'emploi ou par les produits de ces comportements, il y a eu inférence pour les transposer en habiletés ou en qualités nécessaires à ces comportements.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

C H A P I T R E

ÉLABORATION D'INSTRUMENTS DE MESURE Partie II : Développement et implantation

Le chapitre 5 a permis de définir le plus exactement possible le domaine à mesurer par l'instrument. Il reste maintenant à mettre au point un instrument qui mesure adéquatement ce domaine dans le contexte des usages prévus pour cet instrument. C'est l'objet du présent chapitre où se poursuit la présentation des étapes du processus d'élaboration d'instruments de mesure fondés sur la validation de contenu (voir tableau 5.1).

ÉTAPE 4. CONCEPTION DE L'INSTRUMENT DANS SON ENSEMBLE

Le développement de l'instrument de mesure proprement dit débute par des choix relatifs à sa conception générale. Mais, au préalable, il est nécessaire de connaître les **ressources**, les **opportunités** et les

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

contraintes de la situation. Il vaut mieux saisir ces paramètres dès le départ et consacrer son énergie sur ce qu'il est possible d'accomplir à l'intérieur des limites imposées. Ainsi, on doit se préoccuper des budgets, des délais, des ressources humaines disponibles et leur expertise, de la culture de l'organisation et, bien entendu des aspects légaux (chartes des droits, lois du travail, conventions collectives, etc.). Une fois les paramètres de la situation identifiés, on doit concevoir un devis présentant les principales facettes de l'instrument.

FORMAT DE L'INSTRUMENT ET TYPE D'ITEMS

Il faut décider du format de l'instrument de mesure, ou ce que nous avons appelé les conditions d'application (voir figure 2.2). Par exemple, est-ce que l'instrument sera un test ou un examen écrit, une entrevue orale devant un comité, un examen exigeant la manipulation d'un appareil (ordinateur, chariot élévateur, chalumeau à acétylène, etc.), ou encore une mise en situation interpersonnelle dans laquelle le candidat sera appelé à se comporter comme dans la réalité ? Est-ce un instrument collectif auquel on soumet plusieurs candidats à la fois comme dans le cas de la plupart des examens écrits, ou un instrument individuel évaluant une personne à la fois ? Suivant le format retenu, il faut aussi déterminer le **type d'items** ou de stimuli qui seront proposés aux répondants : questions à choix multiples, questions à court développement, analyses de cas et dissertations, résolutions de problème et mises en situation, etc.

Chaque format d'instruments, chaque type d'items comporte ses avantages et ses inconvénients, qu'il convient d'évaluer avec soin à la lumière des finalités établies à l'étape 1 (voir tableau 5.1). La nature du domaine à mesurer joue également un rôle important dans ces choix (Nunnally et Bernstein, 1994). Par exemple, mesurer des connaissances particulières de façon objective chez un grand nombre de candidats peut très bien se faire par un examen écrit composé d'items à choix multiples. En revanche, il vaut mieux procéder oralement, par des questions ouvertes, pour évaluer la capacité de s'exprimer dans une autre langue. La dextérité à utiliser un outil ou une pièce d'équipement pourrait être mesurée par un échantillon de travail au cours duquel diverses tâches seraient demandées au candidat. Pour l'habileté à diriger un groupe de travail, il serait intéressant de procéder à

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

l'aide d'une simulation de groupe avec tâches et rôles assignés aux participants ou simplement avec une entrevue ciblée sur la performance passée du candidat en pareilles situations.

En **règle générale**, le format et le type d'items choisis devraient faire en sorte que la nature de la tâche exigée du candidat à l'instrument se rapproche le plus possible de ce qui se passe réellement dans l'emploi (Murphy et Davidshopher, 1988 ; Mussio et Smith, 1973)[1]. Cependant, d'autres aspects, plus pratiques, peuvent être considérés, comme la facilité d'élaboration de l'instrument, la correction (temps requis et objectivité) et le nombre de candidats à évaluer (Centre de psychologie du personnel, 1984).

DURÉE DE L'INSTRUMENT ET NOMBRE APPROXIMATIF D'ITEMS

La **durée** de passation de l'instrument auprès des candidats et le **nombre approximatif d'items** font partie des aspects à établir. Encore là, il y a des avantages et des inconvénients. Un instrument plus long donne habituellement une meilleure évaluation des candidats en termes de précision (fidélité) et d'étendue (validité de contenu), mais coûte plus cher à utiliser et à corriger. Mais s'il est trop long, il risque d'entraîner la fatigue des candidats et des examinateurs s'il y a lieu (p. ex., lors d'une entrevue, d'une simulation).

MODE DE CORRECTION ET STANDARDISATION

Un des éléments les plus importants qu'il convient de préciser dès cette étape est le mode de correction et son degré de standardisation.

Correction mécanique ou correction clinique. D'une part, la correction peut être standardisée par des règles et procédures strictes, de manière à réduire le jugement et l'interprétation du correcteur. Suivant cette approche **mécanique**, chaque réponse du candidat est cotée en fonction d'une grille (p. ex., grille de correction pour des questions à choix multiples) ou évaluée point par point selon un barème ou un solutionnaire (p. ex., solutionnaire détaillé et chiffré

1. Ce rapprochement entre la tâche exigée par un instrument de mesure et la nature de la performance est abordé de nouveau au cours de l'étape portant sur la création des items et de l'élaboration des conditions d'application, plus précisément lorsque nous traitons du deuxième principe sur la représentativité.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

d'un examen de connaissances constitué d'un répertoire de toutes les réponses possibles et des points accordés). Au terme de la correction, les candidats obtiennent normalement des résultats chiffrés.

D'autre part, la correction peut laisser plus de place à l'interprétation du correcteur ; on parle alors d'une approche plus **clinique** qui repose sur le jugement et la compétence du correcteur. Par exemple, dans le cas de l'entrevue de sélection traditionnelle, l'interviewer analyse et interprète les informations retenues, puis les évalue de son mieux. L'évaluation obtenue par le candidat peut être chiffrée ou il peut tout simplement s'agir d'une description qualitative de ses forces et de ses faiblesses.

On retrouve des modes de correction pour tous les niveaux de standardisation. Ainsi, Huffcutt et Arthur (1994) classent en trois catégories les modes d'évaluation utilisés par les interviewers : 1) une évaluation globale portée sur l'ensemble de la prestation du candidat, 2) une évaluation en fonction de plusieurs critères prédéterminés ou 3) une évaluation de chacune des réponses du candidat en fonction d'un répertoire des éléments attendus.

Avantages et inconvénients. Le grand mérite de la standardisation est la fidélité des résultats obtenus chez les candidats et les avantages qui en découlent. Plus une approche de correction est standardisée, ou **mécanique**, plus elle tend à éliminer les erreurs accidentelles (aléatoires) dues à l'interprétation et au jugement des correcteurs, et plus elle contribue à uniformiser la correction entre les correcteurs (fidélité ou accord interjuges) et d'un candidat à l'autre (fidélité intrajuge). Parce qu'elle contribue directement à la fidélité des mesures obtenues, l'approche mécanique est favorisée en psychométrie (SIOP, 1987). Très objective de nature, c'est une approche que les candidats ont moins tendance à contester et qui est plus facile à défendre lors d'une contestation ou d'un litige. Pour ces raisons, l'approche mécanique est très prisée dans la fonction publique. Toutefois, elle présente des inconvénients. La préparation des outils de correction exige beaucoup de travail, car tous les éléments de réponse attendus doivent être prévus, et comme l'application mécanique de la grille de correction ou du solutionnaire est établie, il y a peu de place pour s'adapter par la suite aux particularités.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Le grand mérite d'une correction plus **clinique** réside dans sa souplesse. Il n'est pas nécessaire de prévoir toutes les réponses possibles : il est possible d'adapter la correction à chaque cas. En contrepartie, plus il y a de place à l'interprétation et au jugement, plus la qualité de l'évaluation repose sur les capacités du correcteur à s'acquitter d'une telle tâche. En contexte de sélection, l'approche clinique est celle que préfèrent les cabinets-conseils pour au moins deux raisons : premièrement, la souplesse de l'approche clinique leur permet de s'adapter rapidement au mandat et à la particularité de la situation ; deuxièmement, il leur serait très difficile et extrêmement complexe d'établir des clés de correction à la fois génériques et standardisés pour toutes les situations rencontrées dans leur pratique.

La fidélité, condition nécessaire mais non suffisante à la validité. La standardisation est un moyen d'augmenter la fidélité, qui est, à son tour, une condition nécessaire pour obtenir la validité. Cependant, il faut rappeler que la fidélité n'est pas une condition suffisante de validité. La standardisation de la procédure de correction ne contribue pas directement à la pertinence des scores accordés par rapport au domaine à mesurer, c'est-à-dire à leur validité. Même si la grille ou le solutionnaire prévoit exactement le nombre de points à accorder, cela ne signifie pas que ce nombre de points reflète l'évaluation que cette même performance aurait vraiment reçue en emploi. Pour être valide, l'outil de correction doit, en plus d'être standardisé, être élaboré de manière à refléter, exactement et sans biais, le domaine de contenu à mesurer (voir étape 6 « Élaboration des outils d'évaluation »).

PROCESSUS D'INTERPRÉTATION ET USAGE DE NORMES

Il est essentiel de connaître la façon dont les évaluations ou les scores obtenus à l'instrument de mesure seront interprétés. Pour l'instant, il faut déterminer si l'on opte pour une méthode relative ou une méthode absolue.

Interprétation normative (en référence à des normes) ou interprétation absolue (en référence au domaine mesuré)[2]***.*** Dans le cas de la méthode normative, la performance de chaque personne est située en fonction de la performance d'un groupe témoin, dont la distribution

2. La méthode en référence au domaine mesuré est aussi appelée approche critériée.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

des résultats constitue ce que l'on appelle des normes. Chaque personne est comparée au groupe témoin et la signification accordée à son résultat est relatif à la performance de ce groupe. Par exemple, ce candidat est le meilleur parmi les 10 candidats à avoir été évalués en entrevue ; ou bien, cette personne fait partie des 25 % supérieurs parmi tous les candidats à avoir répondu à ce test depuis cinq ans.

Toutefois, comparer ainsi le résultat d'une personne à celui des autres ne permet pas de connaître le contenu, la nature de sa performance. Pour cela, il faut recourir à une interprétation **absolue**, qui consiste à décrire la performance de la personne en se référant au domaine de contenu mesuré par l'instrument. Par exemple, ce candidat sait planifier des projets complexes, de plus de 10 millions de dollars, autant dans le secteur privé que public, au Canada et en Asie, etc. Cette candidate maîtrise suffisamment les méthodes de recherche pour faire une recension en bibliothèque et sur Internet. Son niveau d'analyse est tel qu'elle sait faire des résumés de lecture fiables, mais qu'elle est incapable de réaliser des synthèses intégrant des informations complexes et provenant de diverses sources[3].

Implication sur le degré de difficulté des items. Le méthode d'interprétation qui sera choisie se répercute sur la nature des items qui devront composer l'instrument de mesure. Avec une méthode **en référence au contenu**, l'interprétation est absolue ; la position relative de la personne par rapport aux autres n'est pas en cause. Ce qui importe de connaître avec le plus de précision possible, c'est le degré de maîtrise de tous les aspects du domaine de contenu mesurés par l'instrument. Dans ce contexte, la pertinence des items devant faire partie de l'instrument l'emporte sur leur degré de difficulté. Un item, même s'il est relativement facile, doit être inclus dans l'instrument lorsque l'aspect qu'il mesure est un élément important du domaine de contenu (Cronbach, 1990). Par exemple, un examen de compréhension de texte pour des messagers doit permettre de distinguer les candidats capables de comprendre ce qu'ils ont à lire, sans plus. Au-delà de ce niveau, un item n'est plus pertinent. Lorsqu'il s'agit d'outils qui ont pour but de vérifier une compétence de base, il peut même

3. Cronbach (1990) mentionne une troisième méthode d'interprétation, qui renvoie à un critère extérieur à l'instrument de mesure comme la performance en emploi, par exemple. Cette méthode pourra également être appliquée de manière relative ou absolue.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

arriver que l'on ne retienne que des items qui seront éventuellement réussis par presque tous les candidats (Nunnally et Bernstein, 1994). Par exemple, un test de conduite de camion peut avoir pour objectif d'évaluer si les candidats connaissent les règles élémentaires de sécurité routière. Réciproquement, un item très difficile, que presque personne n'arrive à passer, doit être retenu s'il fait partie du domaine à mesurer.

Avec une approche **normative**, l'interprétation repose, non seulement sur les compétences à mesurer, mais aussi sur le niveau de compétences relatif au groupe témoin. Il convient alors de choisir des items dont le degré de difficulté permettra de maximiser la différence entre les individus et de faciliter ainsi la prise de décision. Un item trop facile ou trop difficile n'a pas lieu d'apparaître dans un tel instrument.

Degré de difficulté des items et validité basée sur le contenu. Lorsqu'un instrument de mesure est conçu suivant une démarche basée sur la validité de contenu et visant la représentativité du domaine à mesurer, une approche en référence au contenu est presque toujours implicitement adoptée, du moins pour le choix des items. Cependant, une fois l'instrument élaboré, il est très fréquent de passer en mode normatif pour ce qui est de l'interprétation et de la prise de décision. C'est ce qui se passe dans la plupart des situations de **sélection ou de promotion** du personnel où le nombre de postes disponibles est fixe, peu importe le nombre de candidats. Il est alors naturel de donner la priorité à ceux et celles qui ont obtenu les meilleurs résultats aux outils de sélection (à moins d'un contrat collectif de travail qui fixerait d'autres critères de décision comme, par exemple, l'ancienneté des candidats). Cependant, une utilisation normative d'un instrument fondé sur la validité de contenu sera possible seulement si la distribution des résultats obtenus est suffisamment étendue pour permettre la différenciation entre les candidats et si les différences de résultats obtenus à l'instrument reflètent une réelle différence de rendement en emploi (SIOP, 1987, p. 24 et 32 ; EEOC, 1978, section 1607.14, paragraphe C.9). Par conséquent, on pourra tenter de *maximiser l'étendue des scores en retenant dans l'instrument des items, représentatifs du domaine à mesurer et, si possible, comportant un degré de difficulté qui ne soit ni trop élevé,*

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

ni trop faible. Toutefois, il faut se rappeler que la pertinence de l'item par rapport au domaine de contenu doit primer sur son degré de difficulté[4].

Dans un contexte de **placement** du personnel, **d'évaluation du rendement** ou de **formation**, un processus d'interprétation ou de prise de décision basé seulement en référence au contenu est faisable. Mais, en pratique, il subsiste presque toujours un penchant pour l'approche normative. Il est rare en effet que le niveau requis de compétence ou que les objectifs de rendement n'ont pas été déterminés sans référer aucunement à la performance des autres. La représentativité des items et leur degré de difficulté seront abordés de nouveau (voir étape 5 et étape 8).

NOMBRE DE VERSIONS DE L'INSTRUMENT

Le plus souvent, il n'y a qu'une version de l'instrument. Mais dans certaines circonstances, lorsque les candidats sont appelés à passer plus d'une fois l'instrument de mesure, il importe de pouvoir disposer d'une deuxième forme afin de contrer les effets de l'apprentissage. Composée d'items différents mais équivalents, cette deuxième forme doit être comparable à la première version en ce qui concerne le contenu et le degré de difficulté des items.

ÉTAPE 5. CRÉATION DES ITEMS ET ÉLABORATION DES CONDITIONS D'APPLICATION

Plusieurs étapes préparatoires ayant été complétées, il est maintenant possible de produire l'instrument de mesure et ses diverses composantes : les items (stimuli), les conditions d'application et le processus d'évaluation (voir figure 2.2). La présente étape porte sur les deux premières composantes de l'instrument et la troisième composante sera examinée à l'étape subséquente.

4. Même si l'on a recours à des **notes de passage** fixées en référence au contenu, il faudra bien au terme du processus décisionnel choisir un nombre déterminé de candidat parmi l'ensemble de ceux qui se sont qualifiés. À moins d'y aller d'une façon aléatoire (ou de créer autant de poste que de candidats qualifiés), il est difficile d'imaginer comment on pourra choisir sans comparer les candidats entre eux.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

DEUX PRINCIPES POUR ASSURER LA REPRÉSENTATIVITÉ

Construire un instrument de mesure basé sur la validation de contenu implique que les stimuli et les autres composantes de l'instrument reflètent ce qui se passe en emploi, de manière à mesurer une performance chez les candidats qui soit représentative du domaine ou du sous-domaine de contenu visé (voir SIOP, 1987, p. 23). Pour assurer une telle validité, deux grands principes doivent être considérés.

Principe 1

La répartition des items et des points. Le premier principe, surtout d'ordre quantitatif, a trait à la répartition des items ou des points en fonction des diverses subdivisions du domaine de contenu (voir chapitre 5, section « Structure du domaine à mesurer »). Selon ce principe, *les stimuli (items, questions, problèmes, tâches, etc.) ou les points prévus doivent couvrir* ***entièrement****, dans les* ***mêmes proportions*** *et* ***seulement*** *les aspects du domaines ou du sous-domaine de contenu à mesurer.*

Premièrement, les **stimuli** doivent couvrir **entièrement** tous les aspects (ou les subdivisions) du domaine ou du sous-domaine de contenu (Tziner, Jeanrie et Cusson, 1993). Reprenons l'exemple du directeur d'usine dont le domaine de contenu à mesurer concernait les habiletés en « résolution de problèmes et prise de décisions » et avait été structuré par la combinaison de quatre opérations et de six fonctions de l'entreprise (voir figure 5.4). La validité de contenu sera assurée dans la mesure où les items soumis aux candidats couvriront l'ensemble formé par ces 24 subdivisions du domaine de contenu.

Deuxièmement, les stimuli doivent couvrir les subdivisions dans des **proportions** qui respectent l'importance de chacune d'elles (Gavin, 1977). Par exemple, si la production et la gestion du personnel représentent en importance 75 % de l'emploi d'un directeur d'usine, alors les items (où leur pondération au moment de la correction) doivent respecter cette proportion. Le fait de choisir des items en nombre (ou en importance) égal pour chacune des subdivisions aurait pour effet, par rapport à la réalité de l'emploi, de sous-estimer les fonctions de production et de gestion du personnel au profit des autres fonctions de finance et comptabilité, de marketing, de R-D et de systèmes d'information. Notons que l'importance relative des divers aspects du domaine de contenu peut être basée sur autre chose que la proportion du temps requis ou leur fréquence : par exemple,

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

sur le degré de difficulté des tâches, la gravité des conséquences, de leur relation avec la stratégie de l'organisation, etc. (voir chapitre 5, étape 3, paragraphe « Importance relative des dimensions »).

Troisièmement, il faut que les stimuli soient pertinents **seulement** au domaine ou au sous-domaine à mesurer (Tziner, Jeanrie et Cusson, 1993). Ainsi, les items d'un examen ne doivent pas faire appel à des connaissances que le candidat n'a pas à posséder pour s'acquitter convenablement des tâches reliées à l'emploi postulé (Centre de psychologie du personnel, 1984). Les spécialistes recommandent également que les stimuli ne se rapportent qu'à des aspects **importants** du domaine à évaluer (Centre de psychologie du personnel, 1984 ; SIOP, 1987). Conformément à cette règle, le domaine ou sous-domaine de contenu a déjà été restreint, lors de sa spécification, aux seuls éléments critiques de l'emploi (voir étape 3).

Ce premier principe s'applique aussi à la répartition des **points** prévus au solutionnaire ou dans la grille de correction. Le nombre de points que permettent d'obtenir les parties d'un instrument de mesure ou les instruments d'un même ensemble doit être proportionnel aux aspects du domaine de contenu (Arthur, Doverspike et Barrett, 1996). Par exemple, une décision du Comité d'appel de la Commission de la fonction publique du Canada (1998) a rejeté la validité d'un processus de sélection parce que les points attribués à l'un des instruments de mesure avaient une « importance disproportionnée » par rapport à l'ensemble des exigences de l'emploi[5]. Pour être exact, ce principe s'applique d'abord et avant tout au nombre de points, même si la plupart des auteurs n'en parlent qu'en fonction des stimuli. Comme la validité est un attribut des résultats et non de l'instrument de mesure (voir chapitre 2), une répartition représentative des stimuli ne peut être un gage de validité de contenu. Il faut que la représentativité s'applique aux points (résultat) attribués à chacun de ces stimuli, sinon il n'y aurait qu'apparence de validité.

5. Décision de Pierre Baillie dans l'affaire Laquerre, Breault, Sinh et Ducharme contre Revenu Canada (96-NAR-00153).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Principe 2

La nature de la performance à l'instrument de mesure. Le deuxième principe, d'ordre qualitatif, concerne la nature de la performance du candidat en réponse aux items et aux conditions d'application : *la **performance** exigée et mesurée par l'instrument de mesure doit présenter le plus de similitudes possible avec ce qui est réellement requis dans l'emploi* (voir Gavin, 1977 ; Mussio et Smith, 1973), que ce soit des comportements, des résultats produits ou des connaissances ou des habiletés nécessaires.

Étant donné que la performance et les résultats à l'instrument de mesure découlent des composantes de cet instrument (voir chapitre 2, section « Validité et représentativité des composantes »), il faut faire en sorte que les items et les conditions d'application reflètent le plus possible l'emploi et les situations qui y prévalent (SIOP, 1987). Par exemple, les problèmes d'un examen d'informatique devraient idéalement porter sur des problématiques provenant de l'emploi visé, mettre en application les mêmes logiciels, recourir au même type d'ordinateurs, etc. De plus, le répondant devrait être placé dans les mêmes conditions qu'en emploi, avoir accès aux mêmes ressources, disposer des mêmes délais, etc. Cependant, il est à peu près impossible qu'un instrument de mesure soit en tout point identique à l'emploi ; on n'a qu'à songer à la présentation écrite des problèmes et des réponses, le chronométrage de la passation ou à l'usage de questions à choix multiples (SIOP, 1987). Il faut alors minimiser les différences entre l'instrument et l'emploi de manière à ce que la performance du candidat demeure représentative du domaine de l'emploi. De plus, il faut se rappeler les limites que comporte un instrument trop spécifique à l'emploi (voir chapitre 5, section « Limites à être très spécifique à l'emploi »).

Ce principe concerne aussi le **niveau** de performance à l'instrument de mesure (Centre de psychologie du personnel, 1984 ; Cronbach, 1970 ; Gavin, 1977). Il ne faut pas recourir à des items plus difficiles que dans l'emploi, simplement pour accentuer la différence de performance entre les candidats. Lorsqu'il s'agit de validité de contenu, insiste Cronbach (1990), la discrimination entre les candidats n'est pas l'objectif de l'instrument de mesure. Un tel instrument devrait plutôt servir à vérifier si le candidat atteint un niveau requis de performance, qui correspond alors à une note de passage (SIOP, 1987, p. 24). Par exemple, si tous les candidats à un poste de messager

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

maîtrisent suffisamment la lecture pour occuper cet emploi, il n'est pas approprié de discriminer davantage en utilisant un examen de lecture plus difficile et, ce faisant, donner une chance indue aux candidats mieux éduqués (voir section « Processus d'interprétation et usage de normes »).

RÈGLES POUR LA FORMULATION DES DIRECTIVES ET DES ITEMS

Un instrument de mesure ne peut pas être meilleur que les items qui le composent et rédiger ces items est un art que peu de personnes maîtrisent (Nunnally et Bernstein, 1994). Des règles guidant leur formulation sont exposées dans plusieurs ouvrages (voir Bradburn et Sudman, 1982 ; Cooper et Schindler, 1998 ; Mussio et Smith, 1973 ; Nunnally et Bernstein, 1994 ; Thorndike *et al.*, 1991). Un grand nombre de ces règles peuvent se résumer par le mot clarté : clarté des items et des directives, clarté de la tâche exigée de la part des candidats et clarté de la relation entre chaque item et le domaine de contenu (Nunnally et Bernstein, 1994). De nombreuses règles ont aussi pour but de standardiser l'instrument et son application d'un candidat à l'autre. Bien que ces règles aient été développées dans un contexte d'élaboration d'examens et de tests papier-crayon, plusieurs d'entre elles s'appliquent à tout instrument de mesure, que ce soit une entrevue, une simulation, un échantillon de travail ou autre, et certaines devront être adaptées au contexte de l'instrument envisagé.

Voici quelques-unes de ces règles, principalement tirées ou adaptées du document du Centre de psychologie du personnel (1984) ; certaines font l'objet d'une recommandation dans les *Standards* (1999) ou dans SIOP (1987). Voyons d'abord quelques règles générales qui s'appliquent aux **directives** adressées aux candidats et à **toutes les formes d'items.**

N° 1 Rédiger les items et les directives de la manière la plus simple et la plus succincte possible. La complexité de la formulation grammaticale doit être appropriée aux candidats et à l'emploi concerné.

N° 2 Dans les directives aux candidats, définir la tâche aussi clairement que possible afin que les candidats puissent répondre selon les modalités prévues par l'auteur de l'instrument. Au besoin, il faut prévoir un entraînement ou un échantillon de questions types (voir aussi *Standards*, article 3.20). S'il s'agit

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

d'un instrument de mesure visant à recueillir des comportements (p. ex., par une mise en situation ou un échantillon de travail), indiquer clairement le type de comportements recherché.

N° 3 Dans les directives, préciser la façon d'inscrire les réponses (p. ex., sur le cahier d'examen ou une feuille de réponse distincte) et le temps alloué pour répondre à l'instrument.

N° 4 Informer les candidats de la manière dont les points seront attribués : répartition des points entre les items, application d'une correction négative pour les mauvaises réponses, etc.

N° 5 Indiquer qui aura accès aux résultats.

N° 6 Si les candidats sont évalués par rapport au même contenu, ils doivent répondre aux mêmes questions. Donner aux candidats la possibilité de choisir un échantillon de questions parmi un ensemble entraîne une évaluation de chacun en fonction de standards différents.

N° 7 Chaque item devrait porter sur une idée à la fois. Éviter les items à plusieurs volets.

N° 8 Les items doivent être indépendants les uns des autres afin de permettre à un candidat qui échoue un item de démontrer ses compétences aux autres items. Il n'est pas approprié que la connaissance de la réponse à un item soit une condition pour la réussite d'un autre item, comme dans le cas d'un calcul en chaîne.

N° 9 Ne jamais recourir à la double négation.

Il y a des règles qui s'appliquent plus particulièrement aux **items à développement** (questions ouvertes, analyse de cas, mises en situation).

N° 1 La tâche du candidat doit être clairement définie dans la question. Éviter de poser des questions trop générales ou trop détaillées. Par exemple, la question « Discutez du concept de la comptabilité du prix de revient » est tellement vague que le candidat pourrait répondre à partir d'une multitude d'approches. Par ailleurs, la question « Énumérez trois applications importantes de la comptabilité du prix de revient » est si spécifique que la réponse sera forcément très courte.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

N° 2 Indiquer au candidat l'étendue et l'orientation de la réponse exigée, notamment la longueur, la quantité de détails ou de faits à inclure, etc.

N° 3 Pour chaque item, il devrait être possible de déterminer ce qui constitue une réponse exacte et complète. Si le rédacteur prend la peine d'énumérer les éléments qui doivent entrer dans la réponse, cela lui permettra souvent de déceler les failles importantes dans la question.

N° 4 Utiliser des questions qui exigent des réponses nettement bonnes ou acceptables plutôt que des questions qui ne visent à connaître que le point de vue ou l'attitude du candidat. Cela ne signifie pas qu'il n'y a qu'une seule réponse et que le candidat ne peut pas exprimer son point de vue, mais uniquement que le candidat devrait être évalué sur la démonstration et le raisonnement présentés, plutôt que sur son point de vue.

N° 5 S'il s'agit d'un item à court développement, il devrait être formulé pour qu'une seule bonne réponse soit possible.

Il y a les **items à choix multiples** dont la formulation est particulièrement difficile. Ces items exigent l'application de nombreuses règles.

N° 1 La question de l'item doit être suffisamment claire pour qu'un candidat compétent puisse prédire la bonne réponse immédiatement après sa lecture. Il ne devrait pas être obligé de lire les options proposées (choix de réponse) pour comprendre le sens de l'item.

N° 2 Si la rédaction d'un item repose sur une opinion ou sur une autorité quelconque, il faut indiquer dans la question le nom de l'autorité invoquée ou le titre de la source citée (p. ex., « D'après la théorie des systèmes budgétaires préconisée par Smith, la technique du budget à base zéro est surtout indiquée lorsque… »).

N° 3 Dans la mesure du possible, éviter les questions négatives (p. ex., « Laquelle des caractéristiques suivantes n'est pas une caractéristique essentielle de la technique du budget à base zéro ? »). Les candidats sont habitués à choisir la bonne réponse

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

et risquent de se fourvoyer. Si l'emploi de la tournure négative est jugé nécessaire, on veillera à souligner ou à mettre en relief les mots négatifs.

N° 4 Au moins quatre options de réponse devraient être élaborées pour chaque item, car un nombre inférieur à quatre aide les candidats à deviner la bonne réponse. Cependant, il devient de plus en plus difficile de rédiger des mauvaises réponses valables à mesure que le nombre d'options augmente.

N° 5 Parmi les options proposées, la bonne réponse doit s'imposer d'emblée au candidat compétent. La bonne réponse n'est pas forcément et indiscutablement la seule possible, mais on doit pouvoir démontrer que c'est la meilleure parmi celles qui sont fournies. La bonne réponse doit être correcte pour l'ensemble des circonstances pertinentes que l'on peut raisonnablement rencontrer dans l'emploi concerné (SIOP, 1987, p. 24).

N° 6 Les options de réponse erronées (ou distracteurs) doivent être clairement fausses pour l'ensemble des circonstances pertinentes que l'on peut raisonnablement rencontrer dans l'emploi concerné (SIOP, 1987, p. 24). Il faut prendre soin de ne pas inclure de réponses qui pourraient se révéler correctes dans certaines circonstances.

N° 7 Tous les distracteurs doivent être plausibles. Il est préférable de recourir à des erreurs communes ou à des conceptions erronées que d'inclure des énoncés banals ou invraisemblables. Si aucun candidat ne choisit un distracteur, celui-ci est inutile. Un distracteur non plausible augmente simplement la probabilité de deviner la bonne réponse.

N° 8 La difficulté d'un item à choix multiples est liée étroitement aux options de réponse. Plus le choix de la bonne réponse implique des distinctions subtiles, plus l'item est difficile.

N° 9 Dans les options de réponse, il faut utiliser le moins possible des expressions telles que « toutes ces réponses », « aucune des réponses ci-dessus », etc. Sans ces expressions, un item à choix multiples exige du candidat de trouver, parmi les options présentées, celle qui est la plus vraie. En ajoutant ces expressions comme option, on demande au candidat d'évaluer si chaque (ou toute) option est « assez vraie pour être vraie ». Il est encore

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

moins recommandé d'employer l'option « toutes ces réponses ». Si le candidat constate qu'au moins deux possibilités sont correctes, il saura automatiquement qu'il faut choisir cette dernière option, sans pourtant savoir si la troisième option est aussi correcte.

N° 10 Varier de façon aléatoire l'emplacement de la bonne réponse par rapport aux distracteurs.

N° 11 En rédigeant les options, il faut faire attention à ne pas fournir d'indices qui révéleraient la bonne réponse. Les options ne doivent pas contenir de mots ou d'expressions qui permettent à un candidat astucieux de déceler la bonne réponse sans aucune connaissance de la matière qui fait l'objet du test. Voici quelques exemples d'indices à éviter.

a) Les expressions trop générales (toujours, jamais, seulement, tous, aucun, etc.) peuvent indiquer qu'une option est inexacte.

b) La longueur de la bonne réponse par rapport aux distracteurs est un autre indice : il vaut mieux recourir à des options qui ont toutes relativement la même longueur.

c) Si deux options signifient essentiellement la même chose, aucune d'elles ne peut donc être bonne.

d) Les options qui sont différentes grammaticalement de la question d'un item sont souvent inexactes. Chaque option doit donc s'accorder grammaticalement à la question et former une phrase complète (p. ex., si une question fait appel à une réponse au pluriel, toutes les options doivent être au pluriel).

e) Une option qui chevauche ou en inclut une autre a pour effet de réduire le nombre de possibilités réelles parmi lesquelles le candidat doit choisir la bonne réponse (p. ex., l'option « inférieure à 1,50 mètre » inclut forcément l'option « inférieure à 1,60 mètre »).

Exemples de questions pour l'entrevue de sélection. Afin de venir en aide à ses agents de dotation, le service des ressources humaines d'une organisation a préparé un répertoire de questions d'entrevue de sélection (voir tableau 6.1). Les questions sont classées

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

en fonction des dimensions les plus fréquemment évaluées (autonomie/initiative, relations interpersonnelles, processus de solution de problème, etc.). Pour chaque dimension, trois types de questions sont suggérées : des questions d'auto-appréciation, des questions comportementales et des questions situationnelles.

ÉTAPE 6. ÉLABORATION DES OUTILS D'ÉVALUATION

L'instrument de mesure n'est pas complet sans sa troisième composante : les outils nécessaires à l'évaluation des réponses fournies par les candidats (voir figure 2.2). Parmi ces outils, on retrouve la clé de correction et une procédure pour compiler les scores. Pour certains instruments plus cliniques et plus axés sur le jugement d'examinateurs (p. ex., une entrevue ou une simulation d'une discussion en groupe), ces outils prendront plutôt la forme de grilles d'observation, d'interprétation ou d'évaluation.

Un processus d'évaluation sous le signe de la validité et de l'objectivité. Les outils d'évaluation doivent être conçus et appliqués de telle sorte que les résultats obtenus par les candidats soient représentatifs de leur performance dans le domaine de contenu visé. Il ne sert à rien d'avoir des items ou des conditions d'application représentatives, si les réponses fournies par les candidats sont évaluées par un processus biaisé qui ne reflète pas de façon pertinente le domaine à mesurer. En plus d'être valide, le processus d'évaluation doit être uniforme d'un candidat à l'autre, objectif et impartial ; il en va du traitement équitable des candidats. Pour y arriver, il faut recourir à des procédures et à des outils standardisés, clairs et présentés avec suffisamment de détails de manière à réduire les erreurs d'interprétation chez les examinateurs (voir chapitre 4).

CLÉ DE CORRECTION ET COMPILATION DES SCORES

Une clé de correction est constituée de la liste des **bonnes réponses** (ou des réponses standard ou des éléments de réponse attendus) pour chaque item de l'instrument de mesure, incluant toute variation acceptable s'il y a lieu. Ces éléments de réponse doivent être accompagnés du nombre de **points** prévus. Pour être valide, la clé de correction doit se conformer aux deux principes de la représentativité énoncés plus tôt (voir étape 5). La répartition des points est

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 6.1
EXEMPLES DE QUESTIONS D'ENTREVUE

Dimension à évaluer	Question d'auto-appréciation	Question comportementale	Question situationnelle
Autonomie/ initiative	– Que faites-vous quand vous êtes pris avec un problème administratif que vous ne pouvez résoudre ? (Donnez des exemples.)* – Comment réagissez-vous en période de temps mort ? (Donnez des exemples.) Etc.	– Dans le cadre de vos expériences passées, avez-vous déjà apporté de nouvelles méthodes de travail ? • Si oui, quelle a été votre approche ? • Quelles ont été les principales difficultés rencontrées ? • Comment les avez-vous résolues ? Etc.	– Un client se présente à vous et demande à rencontrer le directeur ; celui-ci est occupé pour le reste de la journée. Que faites-vous ? – Votre supérieur refuse un prêt que vous aviez d'abord recommandé d'accepter. Vous possédez les informations qui prouvent la solvabilité du client. Vous décidez de rencontrer votre supérieur. Quelle sera votre approche ? Etc.
Relations interpersonnelles	– Selon vous, quels sont les avantages et les inconvénients du travail d'équipe ? – Avec quel genre de personnes aimez-vous le mieux travailler ? Etc.	– Parlez-nous d'un problème particulier qu'un de vos employés vous a causé récemment. • Qui était concerné ? • Comment avez-vous traité ce problème ? • Que feriez-vous autrement si une telle situation se représentait la semaine prochaine ? Etc.	– Vous faites face à un surplus de travail important. Vous êtes stressé par la situation et vous demandez de l'aide à un collègue chargé des mêmes fonctions que vous. Ce dernier ne vous accorde pas tout le soutien que vous souhaiteriez. • Quelle sera votre réaction ? • Que lui diriez-vous ? • Que pouvez-vous faire pour obtenir sa collaboration ? Etc.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Processus de solution de conflits	– Quelles sont vos forces et vos faiblesses dans la résolution de conflits ? – Si nous interrogions vos employés sur votre manière de résoudre des conflits, que nous diraient-ils ? Etc.	– Quel est le pire conflit auquel vous avez dû faire face dans votre travail ? • Décrivez-nous la situation. • Qui étaient concernés ? • Qu'avez-vous fait ? • Comment s'est réglée la situation ? Etc.	– Deux employés sont en chicane depuis deux mois. Leur conflit tend à avoir une influence négative sur le climat de votre unité. Le directeur vous demande de remédier à la situation. Que faites-vous ?

* Demander des exemples est une question de type comportemental.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

assujettie à l'obligation de refléter le domaine de contenu, car un trop grand déséquilibre par rapport à ce qui est vraiment exigé dans l'emploi mettrait en péril la validité de contenu. Il faut également que les réponses attendues soient pertinentes à l'emploi concerné et reconnues comme valables par des spécialistes du domaine (*subject matter experts* ou SME). Plus ces spécialistes détiendront les compétences appropriées et plus ils seront nombreux, meilleure sera la démonstration.

Des **directives** précises et détaillées doivent être fournies aux examinateurs sur la façon d'attribuer les points et de calculer les résultats. Habituellement, l'examinateur compare les réponses du candidat avec celles de la clé de correction, puis attribue les points conformément aux directives. S'il est permis d'accorder une partie des points pour une réponse incomplète, les éléments de réponse donnant droit à ces points doivent être clairement établis (Centre de psychologie du personnel, 1984). Une procédure pour les cas litigieux doit également être prévue (attribuer les points en cas de doute, en référer à une autorité compétente, faire reprendre la correction « à l'aveugle », etc.). Des directives claires devraient assurer l'objectivité de la correction, qui se traduit notamment par l'accord de deux (ou plus) examinateurs qualifiés sur les points attribués.

Il est parfois nécessaire de **transformer les résultats** bruts pour les présenter sous une autre échelle, plus facile à comprendre des usagers, plus utile aux preneurs de décision, ou pour uniformiser des résultats provenant de différents instruments ou de parties d'instrument. Il peut s'agir, par exemple, d'une simple transformation en pourcentage, du calcul d'un rang percentile ou de ramener divers résultats sur une échelle de 1 à 10 ; la manière d'effectuer ces transformations doit être décrite en détail (*Standards*, 1999, article 4.1) et ces transformations doivent être fondées à la fois sur le plan logique et psychométrique (SIOP, 1987, p. 31). Si elles ont exigé le recours à des normes (p. ex., rang percentile), il faut décrire et justifier l'usage de ces normes.

OUTILS D'OBSERVATION, D'INTERPRÉTATION OU D'ÉVALUATION

Les recommandations relativement au processus d'évaluation privilégient une approche de type mécanique (voir étape 4, section « Mode de correction et standardisation »). Or, très fréquemment en gestion

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

des ressources humaines, le processus d'évaluation est plus clinique. Le jugement de l'examinateur intervient, à des degrés divers, lors de l'observation, de l'interprétation ou de l'évaluation des réponses des candidats. C'est ce qui se passe notamment pour la plupart des entrevues de sélection, des analyses de cas et autres questions à développement ainsi que pour les simulations dont l'objet est d'étudier le candidat en action. Si tel est le cas, les critères d'évaluation et la procédure de formation des examinateurs doivent être exposés en détail de manière à permettre un haut niveau d'accord entre ces examinateurs. S'il y a usage d'échelles d'évaluation (p. ex., échelle en cinq points de type Likert) ou si des scores sont obtenus par la codification ou la classification des réponses, les outils utilisés et leur application doivent être clairement définis (*Standards*, 1999, article 3.22 à 3.24).

Divers outils peuvent être élaborés pour encadrer le jugement des examinateurs et pour standardiser le processus d'évaluation. Peu importe leur forme, ces outils devront comporter obligatoirement un répertoire des éléments de réponse et de comportement attendus et une échelle d'évaluation.

Éléments de réponses et de comportement attendus. Pour chaque dimension ou capacité à mesurer, il faut dresser la liste des réponses et des comportements typiques d'un candidat qui possède cette dimension ; ces éléments recherchés ou souhaités chez le candidat sont appelés des **indicateurs**. Leur rôle est essentiel : ils contribuent à définir sur le plan opérationnel les dimensions à évaluer en procurant des indices observables, voire quantifiables, de leur présence.

Prenons pour exemple une simulation où le candidat doit analyser un cas, puis faire une présentation orale devant un comité sur sa vision du problème et sur les solutions qu'il propose. Supposons que les dimensions évaluées par cette simulation sont l'analyse, le sens pratique, l'esprit de décision et la présentation orale. Un exemple de répertoire est présenté au tableau 6.2. Dans cet exemple, seule la dimension « présentation orale » est traitée. Cette dimension est d'abord définie sur le plan conceptuel. Par la suite, on retrouve les éléments de comportements attendus ou souhaités chez le candidat. À proprement parler, cependant, ces éléments ne concernent pas que des comportements. En effet, bien que la plupart des éléments soient effectivement des comportements (p. ex., « Dès le début, le plan de

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

l'exposé est présenté », « Les objectifs poursuivis sont rappelés tout au cours de l'exposé », « Le débit est approprié, ni trop rapide, ni trop lent »), certains visent des résultats produits par ces comportements (p. ex., « Les gestes retiennent l'attention », « Le candidat incite l'auditoire à poser des questions et à discuter »).

Tableau 6.2
EXEMPLE D'ÉLÉMENTS DE COMPORTEMENT ATTENDUS POUR LA DIMENSION « PRÉSENTATION ORALE » ÉVALUÉE DANS LE CADRE D'UNE SIMULATION

Dimension évaluée : présentation orale
Définition Capacité à présenter oralement devant un petit groupe des idées de façon claire et structurée, dans un langage approprié et de manière à capter l'attention des auditeurs.
Éléments de comportement attendus (indicateurs) – Dès le début, le plan de l'exposé est présenté. – Les objectifs poursuivis sont rappelés tout au cours de l'exposé. – Tout le sujet annoncé est couvert. – Les transitions d'une partie à l'autre sont faciles pour l'auditoire. – Les idées sont exprimées clairement. – Des exemples pertinents servent à illustrer les idées. – Le vocabulaire et la grammaire sont convenables. – Le débit est approprié, ni trop rapide, ni trop lent. – Les intonations sont nombreuses et stimulantes. – Les gestes retiennent l'attention. – Le candidat incite l'auditoire à poser des questions et à discuter. – Le candidat écoute attentivement les questions et les commentaires de l'auditoire. – Le candidat répond aux questions de l'auditoire. – Etc.

Exemples d'indicateurs pour l'entrevue de sélection. L'entrevue de sélection constitue un instrument de mesure où le jugement des interviewers joue un rôle important dans le processus d'évaluation. C'est un instrument dont la majorité des items (questions) sont à développement libre et pour lequel l'approche clinique est la plus

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

fréquente[6]. Le tableau 6.3 présente un exemple de grille type susceptible d'encadrer les interviewers tout au long du processus d'évaluation ; l'exemple porte sur la dimension « relations interpersonnelles ». Dans cette illustration, le premier groupe d'indices est composé d'exemples de comportements qui sont attendus d'une personne ayant de saines relations interpersonnelles : « Regarde son interlocuteur », « Donne des signes qu'il écoute », etc. Les indices qui apparaissent dans le deuxième groupe font plutôt référence à des résultats produits : « On se sent à l'aise à son contact » et « On aimerait qu'il fasse partie de notre équipe ». Un troisième groupe d'indices porte sur des éléments de réponse recherchés dans les propos du candidat : « Parle des autres en termes positifs ou respectueux », « Prend les moyens pour connaître les besoins et les sentiments des autres », etc. Quant au dernier groupe d'indices, on y retrouve les informations pertinentes à l'expérience et à la formation, susceptibles de renseigner sur les habiletés interpersonnelles développées par le candidat[7].

Une telle grille peut servir d'outil d'observation en cours d'entrevue. Les éléments attendus sont autant d'observations à effectuer ou d'informations à rechercher. L'interviewer peut tenter de mémoriser le contenu de cette grille avant l'entrevue ou simplement la disposer sur son bureau pour consultation durant l'entrevue ; elle peut aussi servir d'outil d'interprétation et d'évaluation. L'entrevue terminée, l'interviewer remplit cette grille à partir des notes qu'il aura prises et de ce qu'il pourra se rappeler, de manière à classer toutes les informations accumulées sur le candidat concernant les relations interpersonnelles. Il restera à l'interviewer à interpréter et à évaluer ces informations.

Des indicateurs observables dans le contexte de l'instrument de mesure. Les éléments attendus, les indicateurs, doivent être observables dans le contexte de l'instrument de mesure. Dans l'exemple de

6. Dans le cas d'une entrevue très structurée dont le mode d'évaluation s'apparente plus à une approche mécanique, l'évaluation serait alors assurée par une clé de correction standardisée comme pour un examen de connaissances ou un test papier-crayon.
7. Dans cet exemple, les éléments de réponses attendus (les indices) sont réunis en fonction des dimensions du domaine de contenu évalué ; les éléments s'appliquent à l'ensemble de l'entrevue, peu importe la question posée. Certains préfèrent regrouper les éléments de réponses attendus pour chacune des questions, les questions étant classées par dimension évaluée. Chaque méthode comporte des avantages et des inconvénients.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 6.3
EXEMPLES D'INDICATEURS POUR LA DIMENSION « RELATIONS INTERPERSONNELLES » ÉVALUÉE DANS LE CADRE D'UNE ENTREVUE DE SÉLECTION

Relations interpersonnelles
Définition
– Entretient un contact agréable et chaleureux avec les autres.
– Agit de façon à ne pas rendre les autres tendus ou mal à l'aise.
– Saisit, cherche à connaître les besoins et les sentiments de son interlocuteur.
Exemples d'indices que la personne possède cette dimension (indicateurs)
– Regarde son interlocuteur.
– Donne des signes qu'il écoute : hochement de tête, « hum hum », acquiescement, etc.
– On a l'impression qu'il est intéressé par nous et ce que l'on dit.
– N'interrompt pas son interlocuteur inutilement.
– Ne cherche pas démesurément à imposer son point de vue.
– Est souriant.
– Fait preuve d'humour, sans être déplacé.
– On se sent à l'aise à son contact.
– On aimerait qu'il fasse partie de notre équipe.
– Parle des autres en termes positifs ou respectueux.
– Parle des autres de façon nuancée, en considérant leurs points de vue, leurs motivations, leurs limites, etc.
– Prend les moyens pour connaître les besoins et les sentiments des autres.
– Lors de différends, recherche des solutions qui respectent l'amour-propre et la dignité des personnes.
– Expérience de travail avec le public, en relation d'aide, en équipe, etc.
– Activités sociales ou sportives exigeant de bonnes relations interpersonnelles.
– Formation relative aux relations interpersonnelles.

grille d'entrevue (voir tableau 6.3), les indicateurs relativement aux relations interpersonnelles sont de nature à se manifester dans le cadre d'une entrevue de sélection si elle est habilement menée par l'interviewer. Tel n'aurait pas été le cas pour les indicateurs suivants :

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

« Écoute ses subordonnés », « Manifeste des marques d'attention envers ses subordonnés (salutation, sourire, événements spéciaux, etc.) » ou « Ses collègues de travail recherchent son contact ». Ces indicateurs peuvent être relatés par les candidats ou bien déduits par l'interviewer à partir de ses observations, mais ils ne peuvent pas être observés directement au cours de l'entrevue. Il s'agit bien des manifestations observables, mais dans un autre contexte comme celui d'une simulation de travail en équipe ou dans le cadre réel du travail.

Voyons de nouveau l'exemple de la simulation où l'on désire évaluer l'habileté à faire un exposé oral (voir tableau 6.2). Tous les indicateurs retenus sont observables lors d'un exposé simulé devant le comité ; mais, si l'instrument de mesure était une entrevue, les éléments ne pourraient plus être les mêmes. Les éléments observables directement lors de la prestation du candidat en entrevue pourraient être maintenus (« Les idées sont exprimées clairement », « Le vocabulaire et la grammaire sont convenables », « Le débit est approprié, ni trop rapide, ni trop lent », etc.), alors que les autres devraient être rejetés (« Dès le début, le plan de l'exposé est présenté », « Les objectifs poursuivis sont rappelés tout au cours de l'exposé », « Le candidat incite l'auditoire à poser des questions et à discuter », etc.). D'autres éléments, qui sont observables en entrevue, pourraient alors s'ajouter. Il pourrait s'agir de produits et de résultats passés (p. ex., « A réussi un ou des cours sur la communication », « A gagné un prix d'art oratoire », « A reçu de bonnes évaluations de la part de ses étudiants »), d'éléments de connaissances mentionnées (p. ex., « Cite les grands principes de la communication devant un groupe », « Énonce les conséquences d'une mauvaise écoute »), de formation (p. ex., « A suivi un cours en exposé oral ») ou d'expérience (p. ex., « A déjà donné des cours à l'université »).

Échelle d'évaluation. L'autre outil essentiel aux examinateurs est une échelle d'évaluation ; elle prend habituellement la forme d'une échelle de quatre à sept gradations. Il est futile de recourir à une échelle dont les gradations seraient plus fines que le degré de précision qu'il est possible d'atteindre avec l'instrument de mesure envisagé ; sa précision ne serait qu'illusoire. De nombreux types d'échelles ont été proposés au cours des années, chacune élaborée selon une méthodologie destinée à leur assurer une plus grande objectivité (voir Guion, 1998, chap. 12). Cependant, les résultats obtenus ne semblent pas avoir convaincu les professionnels de la

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

supériorité de ces échelles, compte tenu des efforts requis pour leur élaboration. Un exemple d'échelle d'évaluation utilisée en pratique apparaît au tableau 6.4 ; son emploi exige une forte dose de jugement de la part des ses utilisateurs. C'est le propre d'une approche clinique dont la fiabilité des résultats est intrinsèquement liée à la compétence et à la volonté des évaluateurs. Par ailleurs, le fait de ne retenir que quatre ou cinq niveaux permet aux évaluateurs de se fixer plus facilement des barèmes, surtout après avoir évalué un certain nombre de candidats.

Tableau 6.4
EXEMPLE D'ÉCHELLE D'ÉVALUATION

Insuffisant : ne répond pas aux exigences du poste.	**Faible :** répond de façon minimale aux exigences, acceptable, à la limite.	**Bien :** satisfait correctement aux exigences du poste pour un rendement adéquat.	**Très bien :** dépasse légèrement les exigences du poste.	**Excellent :** dépasse nettement les exigences du poste.
1	2	3	4	5

ÉTAPE 7. RÉVISION DE LA VERSION EXPÉRIMENTALE PAR DES EXPERTS

Une fois l'instrument élaboré, il doit être revu par des experts afin de vérifier la clarté et la pertinence (ou la représentativité) de son contenu. Pour ce qui est de la **clarté**, on peut demander à des experts d'analyser les items et les directives adressées aux candidats afin de mettre au jour les sources possibles d'ambiguïté, puis de suggérer les correctifs appropriés. Ces experts peuvent être des spécialistes en construction de tests ou en révision de textes. Pour ce qui est de la **pertinence**, il faudra s'en remettre à des experts du contenu à mesurer (*subject matter experts*). A priori, toutes les composantes de l'instrument pourraient être soumises à ces experts pour en juger la représentativité par rapport au domaine de contenu : les conditions d'application, les items et le processus d'évaluation. Si plusieurs experts du contenu ont déjà contribué à l'élaboration de ces composantes, cette étape est redondante.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Pour les items par exemple, il faut s'assurer qu'ils soient représentatifs du domaine de contenu. Après avoir fourni aux experts la définition détaillée du domaine à mesurer, il est courant de leur demander si chacun des items appartient ou non au domaine. Dans le cas où le domaine de contenu a été structuré en plusieurs subdivisions, la tâche des experts consiste plutôt à classer chacun des items par rapport à leur subdivision respective. Le concepteur éliminera les items jugés non pertinents ou dont le classement ne fait pas consensus et ajoutera des items dans les subdivisions où leur représentation est insuffisante (Murphy et Davidshofer, 1988). Rappelons que le degré de difficulté des items doit aussi être jugé pour sa pertinence.

Les experts peuvent travailler seuls ou en groupe. S'il y a plus d'un groupe, il faut maintenir l'uniformité de leur démarche par des directives claires. La présence du concepteur de l'instrument de mesure lors de ces sessions de travail peut être très utile. Par ailleurs, il est essentiel de consigner les données biographiques de chaque expert afin de pouvoir démontrer leur crédibilité en cas de besoin (Mussio et Smith, 1973).

Il n'existe pas à proprement parler de coefficient de validité de contenu. Lawshe (1975) a proposé un « ratio de validité de contenu » (*content validity ratio* ou CVR), qui est en fait une mesure du degré d'accord entre des experts sur la pertinence des items. La méthode consiste à demander à des experts d'indiquer individuellement et pour chaque item, si les connaissances ou les habiletés sous-jacentes sont 1) « essentielles » 2) « utiles mais pas essentielles » ou 3) « non nécessaires » au rendement dans l'emploi concerné (voir Schneider et Schmitt, 1986). Le CVR de chaque item est calculé au moyen de la formule suivante :

Ratio de validité de contenu (CVR) $= \dfrac{n_e - N/2}{N/2}$
où n_e : nombre d'experts qui ont déclaré que l'item est « essentiel » N : nombre total d'experts

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Les items qui n'ont pas été considérés « essentiels » par une proportion donnée d'experts doivent être éliminés. Lawshe propose des valeurs seuils pour évaluer les CVR. Une fois les items retranchés, le CVR moyen peut être calculé pour les items restants. Malgré sa simplicité, la méthode de Lawshe est assez peu utilisée en pratique.

PROCESSUS DE DÉFINITION DU DOMAINE DE CONTENU

L'étape 7 terminée (voir tableau 5.1), une version complète de l'instrument de mesure est achevée, bien que ce soit une version expérimentale qui demande encore quelques ajustements. Voyons le chemin parcouru depuis la spécification du domaine de contenu à mesurer (voir figure 5.2). Les étapes 4 à 7 ont porté sur la construction de l'instrument qui servira à mesurer le domaine de contenu. Il a fallu concevoir des items qui sont autant d'occasions de recueillir auprès des candidats des manifestations observables des diverses dimensions du domaine à mesurer. Suivant la logique de la validation de contenu, les composantes de l'instrument de mesure devraient ressembler le plus possible à ce qui se passe réellement dans l'emploi. Malheureusement, on l'a vu, ce n'est jamais tout à fait possible : les conditions d'application, la nature des items ou même le processus d'évaluation ne peuvent recréer l'emploi tel quel. Cette transformation involontaire de la nature du domaine de contenu est symbolisée dans la figure 5.2 par la texture (ombrage plus pâle) de l'instrument de mesure qui est différente de celle utilisée pour le domaine ou sous-domaine de contenu.

Une autre différence importante par rapport au domaine de contenu distingue l'instrument de mesure : ce dernier n'est qu'un échantillonnage du domaine ou du sous-domaine à mesurer. La plupart du temps, le nombre de stimuli possibles est illimité, alors qu'il faut en restreindre la quantité dans l'instrument. Par exemple, les problèmes de gestion que rencontre un gestionnaire sont innombrables et le contenu de l'instrument n'est forcément qu'un échantillon de tous les problèmes possibles, mais un échantillon représentatif tout de même. Ce phénomène est illustré dans la figure 5.2 par les éléments en gris qui représentent imparfaitement le domaine à mesurer. On voit même que deux éléments excèdent le domaine de contenu, indiquant par là qu'un item peut mesurer partiellement autre chose que le domaine visé. Ce serait le cas, par exemple, d'une question

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

d'entrevue biaisée, telle une mise en situation à l'égard de laquelle certains candidats ayant travaillé dans le service concerné seraient fortement avantagés. Par conséquent, cette question ne mesurerait pas seulement les capacités des candidats à résoudre le problème, mais aussi le fait d'être déjà à l'emploi dans ce service. On peut aussi constater qu'un item en recouvre un autre, comme cela arrive fréquemment dans la réalité où deux problèmes mesurent en partie les mêmes aspects.

À partir de la définition du domaine ou du sous-domaine à mesurer, l'élaboration de l'instrument a requis une bonne dose de jugement de la part du concepteur. Premièrement, lors de la conception initiale, il a fallu présumer que le type d'items retenu et le format de l'instrument dans son ensemble étaient de nature à refléter fidèlement ce qui se passe réellement dans l'emploi. Deuxièmement, choisir les items de manière à constituer un échantillon représentatif du domaine n'a pu se faire sans user de discernement. D'ailleurs, le jugement occupe beaucoup de place dans l'élaboration d'un instrument suivant la logique de la validation de contenu.

ÉTAPE 8. ESSAI DE LA VERSION EXPÉRIMENTALE ET CONTRÔLE DES QUALITÉS MÉTROLOGIQUES

Lorsque c'est possible, l'instrument doit être mis à l'essai auprès d'un ou de plusieurs groupes témoins. L'ampleur de cette expérimentation devra tenir compte de l'usage prévu de l'instrument et de ses conséquences. Les groupes témoins doivent présenter le plus de similitudes possible avec la population cible en ce qui concerne l'étendue et le niveau des compétences mesurées par l'instrument. Par exemple, un examen de connaissances de gestion utilisé à des fins de sélection pourrait être expérimenté auprès d'un échantillon d'employés déjà en poste dans les emplois visés. Il est également souhaitable que le groupe témoin comporte les mêmes caractéristiques démographiques que la population cible (âge, sexe, scolarité, etc.). Enfin, il importe que les conditions d'application lors de l'étude pilote doivent ressembler le plus possible à celles qui prévaudront lorsque l'instrument sera vraiment utilisé (Nunnally et Bernstein, 1994).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

La mise à l'essai sert d'abord à s'assurer que l'instrument fonctionne comme prévu, que les directives et les items sont bien compris par les candidats. Une façon efficace de connaître comment les diverses parties de l'instrument sont comprises est de procéder avec un candidat à la fois et de lui demander de réfléchir à haute voix au fur et à mesure qu'il prend connaissance des directives et qu'il répond aux items. De cette manière, on peut savoir si les candidats interprètent les directives et les items dans le sens désiré lors de la conception de l'instrument. La mise à l'essai permet en outre d'inviter les candidats à donner leur avis sur tout ce qu'ils jugent pertinent pour améliorer l'instrument. Au terme de cette expérimentation, si des correctifs importants sont apportés à l'instrument, on devrait répéter l'essai avec un nouveau groupe témoin et continuer ainsi le cycle jusqu'à l'élimination de toute imperfection. Les outils d'évaluation devraient également être mis à l'épreuve et améliorés si nécessaire.

ANALYSE ET CHOIX DES ITEMS

Même si l'élaboration d'un instrument fondé sur la validité de contenu repose d'abord sur le jugement, une analyse des items basée sur des indices statistiques traditionnels peut être extrêmement utile, sinon essentielle. Une telle analyse fournit des informations sur la manière dont les candidats répondent aux items et sur la façon dont chaque item contribue au score total qu'ils obtiennent à l'instrument (Nunnally et Bernstein, 1994). Ces informations peuvent ensuite aider à sélectionner les items les plus intéressants pour l'instrument, toujours en considérant d'abord la représentativité du domaine de contenu à mesurer (SIOP, 1987, p. 23). Ainsi, dans le cas d'un item à choix multiples, un distracteur qui serait choisi plus souvent que la bonne réponse peut signaler la présence d'une ambiguïté dans cet item. Par contre, un distracteur qui n'est jamais choisi laisse supposer qu'il est facilement détecté comme incorrect. Calculer la proportion de personnes qui réussissent un item (ou le score moyen obtenu à cet item) est une autre donnée utile ; elle indique le degré de difficulté d'un item, niveau qu'il convient ensuite d'interpréter à la lumière des caractéristiques du domaine de contenu. Mais si la majorité des candidats doivent maîtriser cet item et que ce n'est pas le cas, il faut alors chercher à comprendre les raisons de cet écart.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

L'analyse d'items est un champ trop technique pour être traitée adéquatement dans le présent ouvrage et nous recommandons la lecture d'ouvrages spécialisés en psychométrie ou en mesure et évaluation. Soulignons cependant, pour ceux qui sont familiers avec l'analyse d'items, que l'approche peut être très différente pour un instrument basé sur la validation de contenu (Nunnally et Bernstein, 1994 ; SIOP, 1987, p. 23). Ainsi, la corrélation entre chaque item et le score total (soit l'indice de discrimination) n'est pas un critère nécessairement valable pour la sélection des items. Un item peut très bien avoir une corrélation nulle et devoir être tout de même conservé si, par exemple, cette absence de relation ne fait que refléter fidèlement l'indépendance de diverses subdivisions au sein du domaine de contenu. Il en est de même pour l'indice de difficulté qu'il faut utiliser avec beaucoup de circonspection, surtout si le domaine de contenu doit être maîtrisé par la majorité des candidats (cette question a déjà été examinée à l'étape 4, section « Processus d'interprétation et usage de normes », et à l'étape 5, section « Deux principes pour assurer la représentativité »).

QUALITÉS MÉTROLOGIQUES, NORMES ET NOTE DE PASSAGE

Les données recueillies lors des essais peuvent servir à vérifier les **qualités métrologiques** de l'instrument, telles que divers indices de fidélité, la consistance interne ou l'interrelation entre les parties de l'instrument. L'interprétation de tels indices doit être effectuée dans le contexte d'un processus de validation de contenu (Nunnally et Bernstein, 1994). Ces données peuvent aussi être utilisées pour la confection de **normes**, c'est-à-dire une distribution statistique permettant de situer le score d'un individu en relation aux scores d'un groupe témoin représentatif d'une population définie (concept introduit à l'étape 4, à la section « Processus décisionnel et usage de normes »)[8]. La réalisation de normes peut exiger, dans le cas de certains tests, presque autant de travail que la construction du test lui-même. Pour être fiables et représentatives, les normes doivent s'appuyer sur les résultats d'un échantillon suffisamment grand et constitué de personnes semblables à la population cible (*Standards*, 1999, article 4.5). Par exemple, une épreuve du courrier (*In-Basket Test*) de la Commission de la fonction publique du Canada (Centre de

8. Le chapitre 7 est consacré aux notions statistiques de base.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

psychologie du personnel, 1987) a été mise à l'essai auprès de 129 employés, représentant 8 départements et 11 groupes d'emplois ; les normes ont été établies à partir de leurs résultats.

Finalement, si les concepteurs de l'instrument ont opté pour un processus décisionnel en référence à des normes (voir étape 4), les résultats des études pilotes peuvent être très utiles pour fixer la **note de passage**. En effet, à l'instar du processus d'interprétation, il y a deux approches principales pour fixer une note de passage. La première, **en référence à une norme**, consiste justement à faire passer le test à un échantillon représentatif des candidats visés, à compiler leurs résultats, puis à choisir une note en fonction de la proportion de candidats que l'on veut éliminer ou conserver. C'est ce qui a été fait pour une épreuve du courrier employée dans un important ministère, à laquelle ont été soumis, lors d'une étude pilote, 99 superviseurs de premier niveau déjà en poste et jugés représentatifs des titulaires des fonctions visées par les candidats. Après examen des données, les responsables du ministère ont choisi de fixer une note de passage qui éliminerait environ 60 % des candidats.

La deuxième approche vise à déterminer la note de passage directement en fonction du niveau de rendement attendu en emploi, sans égard aux scores obtenus par un groupe témoin. Selon une première variante de cette approche, si l'instrument de mesure a été conçu selon une logique de validation de contenu, alors des experts peuvent établir, par simple comparaison du contenu de l'instrument à celui de l'emploi, le niveau de résultat au test qui correspond au niveau désiré de rendement en emploi (Guion, 1998 ; Maurer et Alexander, 1992). C'est l'approche **en référence au contenu**. Une autre variante existe, où la relation entre les résultats à l'instrument et le rendement en emploi est établie, non plus sur la foi d'experts, mais empiriquement au moyen d'une étude statistique. C'est une approche **en référence à un critère externe**.

Peu importe l'approche utilisée, fixer une note de passage fait toujours appel à un jugement de valeur. En gestion des ressources humaines, c'est l'employeur qui décide du niveau d'excellence de sa main-d'œuvre. La note de passage peut être aussi haute ou aussi basse que les besoins de l'organisation l'exigent, pourvu que l'instrument utilisé soit valide. De plus, même s'il s'agit d'un jugement, il faut être en mesure de fournir un rationnel solide et basé sur des données

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

concrètes (ratio coûts-bénéfices, nombre de postes ouverts, nombre de candidats, politiques de l'organisation, etc.) (SIOP, 1987, p. 32-33). Finalement, la note de passage doit être raisonnable et conforme à des attentes normales d'un rendement acceptable pour ce genre d'employés (Cascio, Alexander et Barrett, 1988 ; EEOC, 1978).

ÉTAPE 9. RÉDACTION DES DOCUMENTS TECHNIQUES

Lorsqu'un instrument de mesure est utilisé pour prendre des décisions relativement au personnel dans une organisation, il faut rendre disponibles les renseignements concernant le développement (ou les données qui justifient l'utilisation) de cet instrument, ainsi que les procédures à suivre pour l'emploi de cet instrument. Les documents portant sur le **développement ou la validation** de l'instrument doivent contenir tous les éléments nécessaires pour qu'un professionnel compétent puisse évaluer la qualité technique de l'instrument et le caractère approprié de son utilisation ; il doit être informé précisément de ce qui a été fait comme démarche par les concepteurs et des résultats qui ont été obtenus. Il va de soi que le rapport doit être objectif et complet, même si des résultats peuvent nuire à la valeur de l'instrument. Quant au manuel de **procédures**, il doit fournir aux usagers les informations leur permettant de maintenir une utilisation appropriée de l'instrument. On doit y préciser les intentions des concepteurs en matière d'usage et les populations cibles, la procédure d'administration, la correction des réponses, l'interprétation des résultats, etc. Tel est l'avis qui se dégage des *Standards* (1999), même lorsqu'il s'agit d'un instrument développé localement pour être utilisé dans une seule organisation. La Society for Industrial and Organizational Psychology (SIOP, 1987) abonde dans le même sens pour tout instrument de sélection. Par ailleurs, on a pu noter que, aux États-Unis, une défense basée sur la validité de contenu d'un instrument peut être efficace devant le tribunal si un tel processus de validation a été bien suivi et documenté avec soin (Arvey et Faley, 1988 ; Schneider et Schmitt, 1986).

Rapport sur le développement de l'instrument. Voilà matière à nous inciter, non seulement à élaborer des outils de mesure avec rigueur et en suivant un processus reconnu pour en assurer la validité, mais aussi à documenter cette démarche, ses justifications et les

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

résultats obtenus. Dans notre cas, la validité de contenu de l'instrument est déjà assurée si les diverses étapes du processus d'élaboration ont été suivies méthodiquement. Il ne reste qu'à rédiger un rapport relatant de façon détaillée la réalisation de chacune des étapes. En principe, la rédaction de ce rapport est déjà passablement avancée. Il s'agit de rassembler les documents produits à chaque étape du processus et à en faire une présentation cohérente.

Il convient de rappeler que certains aspects sont particulièrement importants lors de la démonstration de la validité de contenu. Ces aspects concernent 1) le processus d'analyse de l'emploi, 2) la description de l'emploi, 3) la spécification du domaine de contenu à mesurer, 4) la relation entre l'emploi et le domaine de contenu et 5) la relation entre l'emploi et le contenu de l'instrument, notamment les items ainsi que les éléments de réponse et de comportement attendus. Sachant que le jugement des experts ayant participé aux diverses étapes est au cœur de cette démonstration, leurs compétences et leur expérience par rapport au domaine de contenu doivent être exposées en détail ; il en est de même pour les concepteurs de l'instrument de mesure.

Manuel de procédures. Le manuel de procédures à l'intention des usagers doit consigner toutes les procédures nécessaires à son application et à son usage. On devrait y retrouver 1) les usages et les populations cibles, 2) la procédure d'administration aux candidats, 3) la procédure de correction des réponses et de compilation des résultats, 4) les cadres d'analyses, les normes et autres outils servant à l'interprétation des résultats, 5) les qualifications et la formation de ceux qui peuvent administrer, corriger ou interpréter les résultats. Afin d'assurer l'uniformité, le manuel doit exhorter les utilisateurs à suivre scrupuleusement les procédures d'administration et de correction, au risque de fausser les résultats (SIOP, 1987, p. 30 ; *Standards*, 1999, article 5.1). Si des changements sont apportés, la procédure doit être amendée par écrit, puis être appliquée systématiquement (article 5,2).

ÉTAPE 10. IMPLANTATION ET SUIVI

L'**implantation** de l'instrument de mesure est une opération qui se planifie et s'organise, comme toute intervention dans une organisation. Selon l'ampleur et la complexité de l'implantation, un plan

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

d'action peut être élaboré, incluant les contraintes et les facteurs de risque ; si nécessaire, il faut impliquer les acteurs concernés, employés, cadres, dirigeants, syndicats ou autres associations. Dans le cas d'un nouvel instrument d'évaluation du rendement, par exemple, la gestion de son implantation est susceptible d'être plus importante que les qualités techniques de l'instrument lui-même. Il est indispensable que les personnes en charge de l'application et de l'utilisation de l'instrument soient qualifiées et reçoivent la formation nécessaire pour s'acquitter de leurs responsabilités (SIOP, 1987, p. 33). Pour les outils de sélection, de promotion ou tout autre instrument servant à qualifier une personne, l'organisation a la responsabilité du traitement uniforme des candidats (SIOP, 1987, p. 34). Il faut prévoir des mesures pour que les candidats ne puissent pas obtenir des résultats de manière frauduleuse (plagiat, obtention illicite de renseignements, accès aux bases de données, etc.) [*Standards*, 1999, article 5,6]. L'accès à ces instruments de mesure et à leur matériel doit être strictement contrôlé (SIOP, 1987, p. 34 ; *Standards*, 1999, article 5.7). Les responsables ont le devoir de protéger les droits des candidats au regard du traitement équitable et confidentiel de leurs résultats (*Standards*, 1999, section 8).

Finalement, il ne faut pas négliger, comme c'est trop souvent le cas, de faire le **suivi** de l'instrument et de son application. Des vérifications ponctuelles peuvent être effectuées relativement à tout écart de procédure d'administration de l'instrument, de correction ou d'utilisation qui est faite des résultats. Au besoin, il faut consolider la formation des personnes qui utilisent cet instrument. On devrait également tenir périodiquement des audits afin d'évaluer l'instrument dans son ensemble en fonction des besoins de l'organisation.

En guise de conclusion. Le processus d'élaboration d'instruments de mesure proposé dans cet ouvrage s'appuie sur la logique de la validation de contenu et sur les nombreuses recommandations des experts. C'est un processus exigeant, qu'il convient d'appliquer avec discernement. À chacun, dirigeant, responsable des ressources humaines, employé, syndicaliste ou même magistrat, de juger ce qui doit être mis en pratique dans le contexte particulier de chaque organisation.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

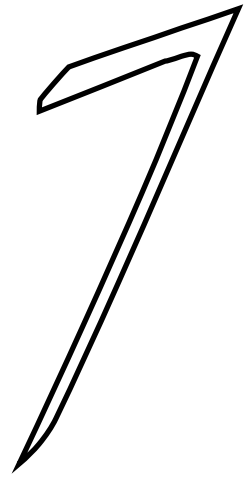
Tous droits de reproduction, de traduction ou d'adaptation réservés

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

C H A P I T R E



FONDEMENTS STATISTIQUES

Ce chapitre porte sur les principaux concepts statistiques utilisés par la psychologie et par les autres sciences du comportement dans l'étude des différences individuelles ; c'est un survol de notions élémentaires, dont plusieurs sont essentielles à la compréhension des concepts faisant l'objet des chapitres précédents. Les notions abordées sont la distribution de fréquences, l'histogramme, la moyenne, l'écart type, le coefficient de variation, la note standard ou note « Z », la distribution normale, le diagramme de dispersion, les coefficients de corrélation et de détermination, ainsi que les régressions simple et multiple, bien que très brièvement. De nombreux exemples illustrent l'utilisation de ces notions en gestion des ressources humaines. Le présent chapitre n'a pas pour but de remplacer un apprentissage spécialisé en statistique, pour lequel il existe déjà d'excellents ouvrages.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

MESURE ET GESTION DES RESSOURCES HUMAINES

Comme pour les traits physiques, tous reconnaissent que les personnes se différencient quant à leurs caractéristiques psychologiques : telle personne éprouve de la facilité à apprendre, telle autre semble très attirée par les activités sociales, une autre est toujours de bonne humeur, certaines sont très indépendantes et acceptent difficilement de se laisser diriger, etc. Les **différences individuelles** sont présentes dans toute la gamme des caractéristiques psychologiques considérées importantes en psychologie du travail : les besoins, les valeurs, les attitudes et les intérêts, les aptitudes, les habiletés et les connaissances, les traits de personnalité ou toute autre dimension du comportement. La gestion efficace d'une organisation passe par la connaissance de ces différences individuelles (Pettersen et Jacob, 1992). Pour bénéficier de tout le potentiel offert par l'ensemble des ressources humaines, les décisions d'ordre stratégique et opérationnel doivent reconnaître ce que chaque personne a de mieux à offrir et ce qu'elle veut offrir. Il ne s'agit pas de renoncer à toute forme de gestion uniforme qui s'appliquerait à un groupe à l'intérieur d'une organisation, mais plutôt de concevoir des politiques et surtout leurs applications de manière à respecter la différence lorsque cela est possible.

Mesurer comporte de nombreux avantages. La mesure des différences individuelles permet d'avoir une approche beaucoup plus rigoureuse en gestion des ressources humaines, particulièrement en ce qui concerne l'évaluation des compétences et la prédiction du rendement au travail. Selon Schneider et Schmitt (1986), cette approche très pragmatique du comportement humain est probablement la contribution la plus importante que les psychologues américains aient faite au monde occidental. Comme pour les autres sciences, la mesure des phénomènes offre de nombreux avantages (Dunnette, 1966 ; Nunnally et Bernstein, 1994).

Premièrement, quantifier une variable selon des règles et des opérations particulières permet une **communication plus précise** entre chacun, spécialiste ou non. Cela élimine les ambiguïtés propres au langage quotidien et contribue à limiter les déformations de la perception individuelle. Qualifier une personne d'excellente en traitement de texte, dénuée d'esprit d'initiative ou de très douée en négociation est flou pour l'interlocuteur et peut donner lieu à bien

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

des interprétations ; mais, mentionner que sa performance au traitement de texte est de 75 mots à la minute avec en moyenne 2,4 fautes lève toute ambiguïté.

Deuxièmement, la quantification **contraint à préciser** ce dont on veut parler, éliminant ainsi les définitions vagues. En effet, il est impossible de quantifier l'initiative ou la négociation sans d'abord définir ces expressions. Se prêter à l'exercice nous convainc assez rapidement de l'imprécision de ces termes et, par conséquent, de leur propension à semer la confusion. Comment alors considérer rigoureusement de telles variables lors d'une opération de sélection, de promotion ou même d'évaluation du rendement ? La quantification est l'antidote des étiquettes ambiguës. Un phénomène qui ne peut être quantifié est souvent un phénomène qui ne peut être défini précisément. En conséquence, il sera difficile de l'intégrer à un processus de gestion que l'on veut logique et rationnel. Certains seront tentés de soulever la question de la place de l'intuition dans la gestion. Mais, au fait, qu'entend-on par intuition ? Sur quoi est fondée cette interrogation ? Ce débat ne cadre malheureusement pas dans les objectifs du présent ouvrage.

Troisièmement, la mesure autorise **l'observation objective** des phénomènes et ouvre la voie à l'accumulation des données empiriques : c'est l'approche scientifique. Le développement du savoir en sciences exige que la logique soit systématiquement confrontée à la réalité, que les intuitions soient corroborées par les observations. Aux intuitions et à la découverte logique, on ajoute l'expérience et la connaissance par la démonstration empirique. L'expérience a corroboré, par exemple, que l'entrevue de sélection peut être doublement efficace si elle est structurée et porte sur les aspects précis de l'emploi à pourvoir. Certains visionnaires en avaient eu l'intuition ; maintenant, c'est un phénomène vérifié. Cependant, toutes les intuitions ne sont pas toujours confirmées par l'expérience. Ainsi, certains visionnaires ont prétendu que la graphologie (étude des individus par leur écriture manuscrite) était un moyen efficace de prédire le rendement au travail ; jusqu'à ce jour, l'expérience n'a pas soutenu cette prétention et cela ne demeure encore qu'une intuition.

La rigueur concerne aussi la gestion des ressources humaines. La mesure des différences individuelles est indéniablement une incitation à la rigueur. Malgré ces avantages certains, il ne manquera pas

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

de gestionnaires pour objecter que la pratique a des impératifs d'efficacité qui s'accommodent très mal de toute cette lourdeur scientifique et qu'il vaut mieux laisser la science aux chercheurs. Il est en effet exact qu'il n'est pas toujours possible, ni même souhaitable, de recourir à des méthodes trop exigeantes du point de vue de la rigueur. En fait, il tombe sous le sens que *les moyens mis en œuvre pour arriver à prendre une décision ou à résoudre un problème doivent tenir compte des délais dont on dispose et surtout avoir une ampleur proportionnelle à l'importance de la situation et de ses conséquences*. Il revient donc au gestionnaire d'évaluer les coûts ou les inconvénients de son incertitude et de doser le degré de rigueur en conséquence, tout en sachant qu'il ne parviendra jamais à se libérer complètement de son incertitude. De plus, il doit se rappeler que les efforts additionnels au niveau de la rigueur ont probablement un rendement décroissant au regard de la diminution de l'incertitude, de sorte que le point mort est vite atteint lorsqu'il s'agit de gestion courante. Néanmoins, toutes ces contraintes pratiques ne changeront rien au fait que la valeur d'une conclusion ou d'une décision ne peut être supérieure à celle des données sur lesquelles elle s'appuie (DeVellis, 1991). Il y a donc de nombreuses circonstances où le gestionnaire de personnel serait plus avisé d'adopter une démarche rigoureuse et fondée sur des données objectives.

MESURE DES DIFFÉRENCES INDIVIDUELLES

Chaque individu, objet ou événement comporte des particularités qui le distinguent des autres ; ces distinctions, attributs, ou différences individuelles dans le cas des personnes sont appelés des **caractères** ou des **variables**. La taille, le sexe, la couleur des yeux, la nationalité, la force physique, l'instruction, l'intelligence et la détermination sont des exemples de caractères s'appliquant à des **personnes**. La couleur, la taille, le poids, la texture, etc., sont des caractères attribuables aux **objets** de nature physique, alors que la durée, l'époque, l'ampleur, etc., concernent plutôt les **événements**. Une organisation, par exemple, qui est un objet en partie physique, en partie conceptuel, se définit par des caractères comme l'âge, le nombre d'employés, le volume des ventes, le secteur d'activités, la localisation, l'immobilisation, etc.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

ÉCHELLES DE MESURE

Il y a quatre façons de mesurer les individus, quatre ensembles distincts de règles permettant d'attribuer des valeurs numériques aux différences individuelles (Catano *et al.*, 1997). Désignées par l'expression « échelles de mesure », ces quatre façons renvoient aux types suivants : échelle nominale, échelle ordinale, échelle d'intervalle et échelle de rapport. Le type d'échelle de mesure employé déterminera le genre de traitement statistique qui pourra être appliqué à ces mesures.

Échelles nominales. Un caractère peut être exprimé de manière quantitative ou qualitative, selon qu'il est mesuré ou non (Baillargeon, 1989). Dans le cas de caractères **qualitatifs**, les **modalités** que peut prendre un caractère sont simplement des catégories permettant de classer les personnes, les objets ou les événements. Ces échelles de classification constituent le niveau de mesure le plus élémentaire : ce sont des échelles nominales (voir tableau 7.1). Les personnes sont classées en fonction de leurs traits physiques (sexe masculin, couleur des yeux, etc.), en fonction de leurs caractéristiques psychologiques ou comportementales (intelligent, fonceur, bilingue, etc.) et parfois en fonction de leurs statuts, de jugements portés par la société ou de tout autre attribut (marié, de nationalité canadienne, riche, célèbre, de formation en psychologie, etc.).

Les modalités constituent des étiquettes ne comportant intrinsèquement aucune information quantitative, donc aucune relation d'ordre entre elles : le sexe masculin n'est ni mieux ni pire que le sexe féminin. S'il y a une relation d'ordre entre les modalités, elle provient forcément d'un cadre de référence extérieur aux étiquettes mêmes. Par exemple, le fait d'être marié ou d'avoir des enfants n'a de mérite qu'en fonction de certaines valeurs sociales, morales ou spirituelles. Les chiffres peuvent servir à exprimer les modalités d'un caractère nominal, sans pour autant en faire une échelle quantitative. Le chiffre prend alors valeur d'étiquette, sans signifier d'ordre entre les personnes, comme par exemple les numéros sur les chandails des joueurs de hockey ou les numéros assignés aux employés par le service du personnel. Les modalités traduisent entre elles le concept « différent de ».

Une échelle nominale, quoique élémentaire, comporte certaines exigences, communes d'ailleurs à tous les autres niveaux de mesure. Les modalités doivent être **mutuellement exclusives**, afin que toutes les personnes classées dans une même catégorie ou portant la même

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 7.1
ÉCHELLES DE MESURE

Échelle de mesure	Définition	Exemples
A) Qualitative		
Nominale	Les modalités (ou valeurs) sont des étiquettes qui permettent d'établir un **classement**, mais sans établir de relation d'ordre entre les classes. Les modalités véhiculent le concept « différent de ».	Sexe État civil Nature des études Langue maternelle
B) Quantitative		
Ordinale	Les modalités sont des valeurs ordonnées suivant une **mise en rang**, mais qui n'indiquent pas le degré relatif de différence entre elles. Les modalités véhiculent les concepts « plus grand que », « plus petit que » et « égal à ».	Le meilleur employé du mois Candidats « à voir », « en attente » ou « à rejeter ».
Intervalle	Les modalités sont des valeurs qui indiquent le degré relatif de **différence entre elles**. Les écarts qui séparent les modalités adjacentes sont égaux.	Les scores à la plupart des tests psychométriques, des examens de connaissances ou de compétences.
Rapport	Les modalités sont des valeurs qui indiquent le degré de **différence par rapport à un point zéro.** Le rapport entre deux modalités peut être calculé (p. ex., 25/50 = 0,5).	Âge Taille Force exercée sur un levier Nombre d'unités produites Volume des ventes Nombre de jours d'absence

étiquette aient quelque chose en commun et qu'elles se distinguent, par ce caractère, des personnes portant une étiquette différente. Les modalités doivent aussi être **exhaustives**, c'est-à-dire qu'elles englobent tous les cas possibles relativement à ce caractère. Par exemple, « marié » et « célibataire » sont des modalités de l'état civil qui sont mutuellement exclusives : une personne ne peut être à la fois mariée et célibataire. En revanche, ces deux modalités ne sont pas exhaustives : il y a des personnes qui vivent en union libre, qui sont séparées ou

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

divorcées. Cependant, des modalités comme « en union libre », « divorcé » et « marié » ne seraient pas mutuellement exclusives, car une personne peut cumuler plus d'un tel état civil à la fois. Élaborer l'échelle nominale permettant de classer les personnes selon leur état civil n'est plus une mince affaire !

En gestion des ressources humaines, les échelles nominales sont constituées le plus souvent par des données de type biographique (sexe, état civil, nature des études, type d'emploi, langues parlées, etc.). Ces données, recueillies par une formule quelconque ou lors d'une entrevue, peuvent servir à la sélection des candidats, à la promotion, à l'implantation de programmes d'égalité en emploi ou à la gestion de plans d'avantages sociaux ou d'autres programmes du service du personnel.

Échelles ordinales. Plusieurs des caractéristiques humaines peuvent engendrer des données **quantitatives**. Les modalités que prennent ces caractéristiques peuvent être exprimées par des chiffres qui indiquent un certain ordre entre les modalités. Il existe trois niveaux de mesure quantitative ; ces trois niveaux vont du plus élémentaire au plus complet, de sorte que le niveau suivant possède les caractéristiques du niveau qui le précède (Sarrazin et Lussier, 1993). Le premier niveau est constitué des échelles ordinales (voir tableau 7.1). Les modalités que peut prendre un caractère de niveau ordinal sont des chiffres qui représentent en fait une mise en rang. Les personnes classées dans la catégorie 1 ont quelque chose de plus, concernant la caractéristique mesurée, que toutes les personnes classées dans les autres catégories ; les personnes classées dans la catégorie 2 ont quelque chose de plus que toutes les personnes classées dans les autres catégories, à l'exception de celles de la catégorie 1, et ainsi de suite. Tous les systèmes de mise en rang sont des mesures de niveau ordinal : rang scolaire, percentile, médailles olympiques, etc. L'échelle ordinale ajoute le concept « plus grand que » au concept « différent de » de l'échelle nominale (Dunnette, 1966). En gestion des ressources humaines, on retrouve quelques exemples de ces échelles : le meilleur employé du mois, le classement des vendeurs en fonction de leur volume de ventes, le rangement des candidats en trois catégories (p. ex., « à voir », « en attente » et « à rejeter ») après avoir analysé leurs curriculum vitæ, etc.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Échelles d'intervalle. Le niveau de mesure suivant est l'échelle d'intervalle. Dans l'échelle ordinale, le degré relatif de différence entre les modalités n'est pas connu ; seule l'ordre des modalités l'est. Par exemple, un coureur olympique peut être premier avec quelques millièmes de seconde d'avance sur le deuxième, qui, lui, devance son plus proche rival de plusieurs dixièmes. La couleur des médailles olympiques indique l'ordre d'arrivée des coureurs, mais non l'écart qui les sépare. L'échelle d'intervalle indique, en plus de l'ordre des modalités, les écarts qui les séparent. En fait, l'écart entre les modalités adjacentes est maintenu égal. L'échelle de température Celsius est un exemple d'échelle d'intervalle ; le changement de un degré est le même tout au long de l'échelle de température, de sorte que l'écart de température entre 20 et 25 est le même que celui entre 50 et 55 pour ce qui est du mouvement de la colonne de mercure.

Les résultats à la plupart des tests psychométriques, développés conformément aux règles reconnues par la psychométrie, peuvent être considérés de niveau d'intervalle. Même si, à proprement parler, on ne peut certifier que l'écart est parfaitement constant entre tous les points de l'échelle, plusieurs considérations théoriques et pratiques militent en faveur d'une telle pratique (Dunnette, 1966 ; Schneider et Schmitt, 1986). Il en est de même pour plusieurs caractères qui, à première vue, semblent de niveau d'intervalle, mais pour lesquels la constance des écarts n'est pas assurée. Prenons l'exemple du nombre d'années d'études. Pouvons-nous présumer que l'apprentissage effectué en première année du primaire est de même ampleur et de même nature que celui de la quatrième année du secondaire ? Malgré une réponse inévitablement nuancée à cette question, il est plus pratique de continuer à considérer ces caractères comme étant de niveau d'intervalle.

Échelles de rapport. Le dernier niveau est l'échelle de rapport, dont le zéro représente l'origine absolue, comme dans zéro kilo ou zéro seconde. Non seulement l'écart entre les modalités est connu, mais il est désormais possible d'établir des rapports entre elles (p. ex., deux fois plus lourd ou la moitié moins rapide). L'âge, la taille, la force exercée sur un levier, le nombre d'opérations effectuées par unité de temps, le volume des ventes, le nombre de jours d'absence sont autant d'exemples de caractères du niveau de l'échelle de rapport. Les tests psychométriques n'ont pas la prétention de relever de ce niveau de mesure, parce que le zéro absolu n'est pas quelque chose de connu. Par exemple, à quoi pourrait ressembler une personne ayant zéro

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

d'intelligence, zéro en coordination visuomotrice ou zéro en créativité ? On ne peut répondre à cette question, parce qu'il est difficile de définir un niveau zéro à l'égard d'un trait humain.

Résumé. Les chiffres qui expriment les modalités d'un caractère peuvent avoir différentes significations selon l'échelle de mesure à laquelle ils appartiennent. Par exemple, le chiffre 7 peut avoir plusieurs significations : soit un numéro d'employé utilisé comme simple étiquette (échelle nominale), soit le septième sur une liste de candidats qui en comprend 12 (échelle ordinale), soit une note de 7 sur 10 lors d'un examen (échelle d'intervalle) ou encore 7 pièces assemblées dans une journée (échelle de rapport).

Échelles de mesure et traitements statistiques. Le niveau de mesure d'une caractéristique détermine les calculs et les traitements statistiques qui peuvent être appliqués. Par exemple, il est insensé de vouloir calculer une moyenne pour les modalités d'une caractéristique de niveau nominal (p. ex., 3 hommes et 4 femmes ne peuvent donner une moyenne de 3,5 pour le caractère sexe). Par contre, il est possible de dénombrer les personnes de chaque sexe ou d'identifier la catégorie qui comprend le plus de personnes. Une échelle ordinale ne permet pas non plus de calculer une moyenne, parce que les diverses modalités, qui sont des rangs, ne sont pas des valeurs séparées par un même écart ; une première et une troisième place ne sont pas équivalentes à deux deuxièmes places. Pour le calcul d'une moyenne, il faut un niveau de mesure d'intervalle ou de rapport, comme dans le cas de l'âge d'un groupe ou des résultats à des examens (Baillargeon et Martin, 1994). De manière générale, les méthodes statistiques qui s'appliquent aux échelles d'intervalle et de rapport sont les mêmes. On les appelle **paramétriques**, par opposition aux méthodes non **paramétriques** qui traitent les échelles des niveaux nominal et ordinal[1].

Les notions statistiques qui sont présentées dans ce chapitre sont des notions paramétriques pour la plupart. Seul ce niveau de traitement a été retenu parce que, d'une part, la majorité des caractéristiques

1. Il est possible de considérer les modalités d'un niveau de mesure comme appartenant à un niveau inférieur. Par exemple, des mesures d'intervalle peuvent être traitées comme des mesures ordinales ou même nominales. Cette substitution, parfois pratique, entraîne cependant une perte d'information. Par contre, le passage vers un niveau supérieur n'est pas acceptable.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

humaines considérées en gestion du personnel sont mesurées à ces niveaux ou considérées comme telles et que, d'autre part, ce niveau de mesure constitue en quelque sorte le prototype classique servant à expliquer les cadres de référence faisant l'objet du présent travail.

NOTES BRUTES, DISTRIBUTION DE FRÉQUENCES ET HISTOGRAMME

Notes brutes. D'innombrables caractéristiques physiques, psychologiques ou comportementales peuvent être mesurées. Les notes brutes sont simplement les valeurs attribuées aux diverses modalités de ces caractéristiques (p. ex., 1 m 64, 64 kg ou 85 % à un test de connaissances). Il arrive souvent, surtout avec des variables psychologiques et comportementales, que les notes brutes n'aient pas grande signification en soi. Prenons l'exemple suivant des résultats à un examen d'une classe de 30 étudiants à un cours universitaire (voir tableau 7.2). Un étudiant qui reçoit sa note à un tel examen, disons 24 sur 30, ne sait pas trop quoi penser de sa performance. Pour mieux saisir l'ensemble des résultats, les notes ont été rangées par **ordre croissant** ; cela permet de voir l'éventail des notes de la plus basse à la plus forte, leur étalement relatif et les notes les plus fréquentes. Cependant, dès que le nombre de notes devient important, un tel rangement, même s'il demeure utile, demande de tenir compte de trop de données en même temps : un résumé statistique s'impose.

Distribution de fréquences. La première étape d'une analyse de données consiste à en effectuer une présentation succincte et intelligible (Baillargeon, 1989). La distribution de fréquence et l'histogramme permettent de regrouper les données semblables, ce qui facilite la visualisation d'un grand nombre de données. La distribution de fréquences est la compilation de la fréquence des observations **par modalité**, s'il s'agit d'une variable **discontinue** (variable quantitative ne pouvant prendre qu'un nombre limité de valeurs, comme le nombre de subordonnés), ou **par classe**, s'il s'agit d'une variable **continue** (variable quantitative pouvant prendre toutes les valeurs d'un intervalle, comme les scores à un examen ou l'ancienneté accumulée dans une organisation) [Baillargeon, 1989]. Pour illustrer une variable continue (voir tableau 7.3), nous avons utilisé les données du tableau 7.2.

Histogramme. L'histogramme est une représentation graphique d'une distribution de fréquences. Dans le cas d'une variable **discontinue**, il est constitué d'autant de bâtons que de modalités et la

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 7.2
NOTES BRUTES, NOTES DE DÉVIATION ET NOTES STANDARD

Sujet	Note brute (X_i)	Note de déviation ($X_i - M_x$)	Note de déviation élevée au carré	Note standard (Z_i)
1. Pierre	13	–8,33	69,39	–2,26
2. Nancy	15	–6,33	40,07	–1,72
3. Micheline	16	–5,33	28,41	–1,44
4. Charles	17	–4,33	18,75	–1,17
5. François	17	–4,33	18,75	–1,17
6. Sylvie	17	–4,33	18,75	–1,17
7. Bernard	18	–3,33	11,09	–0,90
8. Donald	19	–2,33	5,43	–0,63
9. Gilles	19	–2,33	5,43	–0,63
10. Gérald	19	–2,33	5,43	–0,63
11. Mario	20	–1,33	1,77	–0,36
12. Anne	20	–1,33	1,77	–0,36
13. René	20	–1,33	1,77	–0,36
14. Denis	21	–0,33	0,11	–0,09
15. Daniel	22	0,67	0,45	0,18
16. Lyne	22	0,67	0,45	0,18
17. Stéphane	22	0,67	0,45	0,18
18. Nathalie	23	1,67	2,79	0,45
19. Luc	23	1,67	2,79	0,45
20. Marie	24	2,67	7,13	0,72
21. France	24	2,67	7,13	0,72
22. Alain	24	2,67	7,13	0,72
23. Danielle	24	2,67	7,13	0,72
24. Paul	25	3,67	13,47	0,99
25. Mireille	25	3,67	13,47	0,99
26. Johanne	26	4,67	21,81	1,27
27. Linda	26	4,67	21,81	1,27
28. Serge	26	4,67	21,81	1,27
29. Manon	26	4,67	21,81	1,27
30. Robert	27	5,67	32,15	1,54
SOMME (Σ) =	640	0,10	408,67	0,03

Moyenne des X_i (M_x) $= \frac{\Sigma X_i}{N} = \frac{640}{30} = 21,33$ Écart type des X_i (S_x) $= \sqrt{V_x} = \sqrt{13,62} = 3,69$

Variance des X_i (V_x) $= \frac{\Sigma(X_i - M_x)^2}{N} = \frac{408,67}{30} = 13,62$ Note standard (Z_i) $= \frac{X_i - M_x}{S_x} = \frac{13 - 21,33}{3,69} = -2,26$

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 7.3
DÉPOUILLEMENT ET DISTRIBUTION DE FRÉQUENCES

Classe*	Dépouillement	Fréquence
moins de 13		0
$13 \leq X < 15$	✓	1
$15 \leq X < 17$	✓✓	2
$17 \leq X < 19$	✓✓✓✓	4
$19 \leq X < 21$	✓✓✓✓✓✓	6
$21 \leq X < 23$	✓✓✓✓	4
$23 \leq X < 25$	✓✓✓✓✓✓	6
$25 \leq X < 27$	✓✓✓✓✓✓	6
$27 \leq X < 29$	✓	1
29 et plus		0

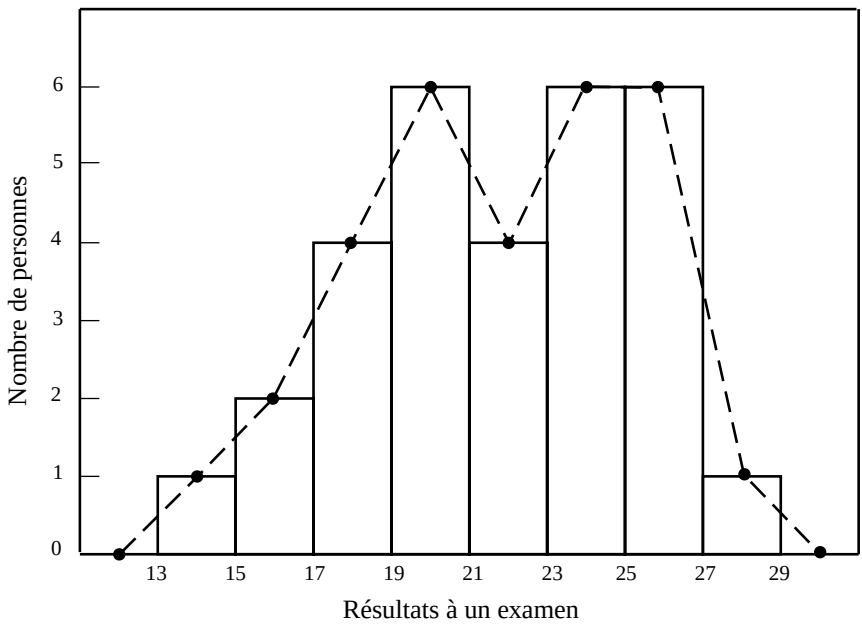
* La détermination du nombre de classes et de leur amplitude a été faite à l'aide de la formule de Sturges (Baillargeon, 1989).

longueur de chaque bâton est proportionnelle à la fréquence de la modalité correspondante. Dans le cas d'une variable **continue**, il est constitué de rectangles dont la base est égale à l'intervalle de chaque classe et la hauteur, proportionnelle à la fréquence. À partir de la distribution de fréquences, l'**histogramme** présenté ci-dessous a été compilé (voir figure 7.1, trait plein). Pour représenter l'allure générale de la distribution à l'aide d'une courbe, on peut recourir au **polygone de fréquences** (Baillargeon, 1989), qui est obtenu en reliant par des segments de droites le centre du sommet de chaque rectangle de l'histogramme (voir figure 7.1, trait pointillé).

L'histogramme est très utile, car il permet de visualiser la distribution d'une série de données. Par exemple, quels sont les résultats que les étudiants obtiennent au concours de l'Ordre des comptables agréés ? Le tableau 7.7 (voir troisième colonne) rapporte ces résultats pour un groupe de 46 candidats ; ces résultats sont corrigés sur 400 points et la note de passage est 240. La compilation de l'histogramme (voir figure 7.2) facilite l'analyse des résultats. On constate que les résultats varient de 150 à 370 et qu'ils se répartissent également de part et d'autre de la note de passage, révélant un taux de réussite d'environ 50 %. Les résultats autour de la note de passage sont nombreux, alors qu'ils diminuent à mesure que l'on s'en éloigne.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 7.1
HISTOGRAMME ET POLYGONE DE FRÉQUENCES



MOYENNE (M)

La mise en rang des notes brutes, la distribution de fréquences et l'histogramme constituent des représentations sous forme de tableaux ou de graphiques servant à l'examen d'une série de données. Il est également possible de représenter une série de données par des indices numériques qui pourront résumer d'une certaine façon l'ensemble des données. La moyenne et l'écart type sont deux de ces indices les plus utilisés à cette fin. La moyenne s'obtient par la somme des notes, divisée par le nombre de personnes qui ont reçu une note ; on peut la calculer grâce à la formule suivante :

$$\textbf{Moyenne (M)} = \frac{\Sigma X_i}{N}$$

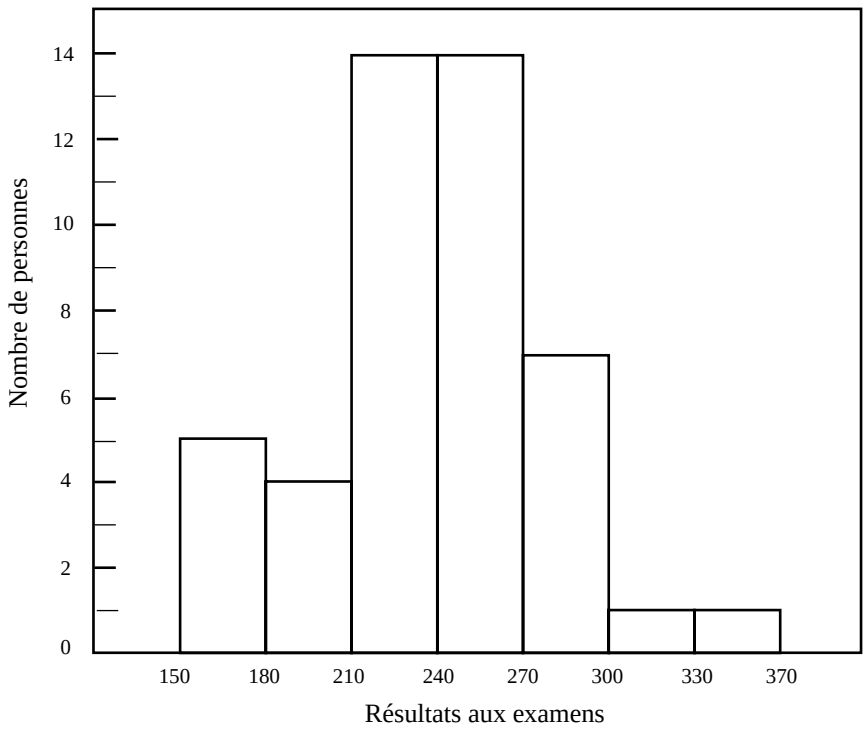
où Σ : somme de tous les X_i

X_i : chaque observation (note, résultat ou autre mesure)

N : nombre de X_i

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 7.2
HISTOGRAMME DES RÉSULTATS DE 46 CANDIDATS AUX EXAMENS DE L'ORDRE DES COMPTABLES AGRÉÉS



En recevant leur note, les étudiants s'empressent habituellement de demander quelle est la moyenne de la classe. La moyenne des 30 étudiants à l'examen est de 21,33; les calculs relatifs sont rapportés au tableau 7.2. La moyenne est un indice de la tendance centrale du groupe : c'est la note autour de laquelle les autres notes se distribuent de part et d'autre.

NOTES DE DÉVIATION, VARIANCE (V), ÉCART TYPE (S) ET COEFFICIENT DE VARIATION (CV)

Note de déviation. Une fois que la moyenne de la classe est connue, un étudiant pourra comparer son résultat et se situer par rapport à l'ensemble du groupe. Il pourra compter le nombre de notes qui le

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

sépare en plus ou en moins de la moyenne : c'est sa note de déviation. Par exemple, la note de déviation de Bernard (sujet n° 7) est de –3,33, soit sa note brute 18, moins la moyenne 21,33 (voir tableau 7.2). Inquiet, Bernard est sans doute intéressé à connaître aussi la dispersion des notes. En effet, il sait, du moins intuitivement, qu'avoir 3,33 sous la moyenne est moins pire si les notes varient de plus ou moins 8 points par rapport à la moyenne que si elles varient de seulement plus ou moins 3 points. L'ensemble des notes de déviation indique ce degré de dispersion ou d'étalement des données. Si les notes de déviation se répartissent étroitement autour de la valeur zéro, c'est parce que les notes brutes sont très proches de la moyenne, et si les notes de déviation sont importantes, les notes brutes sont plus étalées (Dunnette, 1966).

Variance (V). Les notes de déviation peuvent être résumées en un seul indice : l'écart type. C'est l'indice de dispersion le plus utilisé en statistique pour les mesures de niveau d'intervalle et de rapport. L'écart type s'obtient en calculant d'abord la variance, qui est la moyenne des notes de déviation au carré. La formule de la variance est la suivante[2] :

$$\textbf{Variance (V)} = \frac{\Sigma(X_i - M_x)^2}{N}$$

où X_i : chaque observation (personne, objet ou événement)

M_x : moyenne des X_i

Σ : somme de tous les $(X_i - M_x)^2$

N : nombre de X_i

La variance est de 13,62 pour l'ensemble des notes brutes des 30 étudiants ayant passé l'examen. Les calculs sont rapportés au tableau 7.2.

2. Si les données forment un échantillon et que l'on veut obtenir l'estimation de la variance de la population qu'elles représentent, le dénominateur N est remplacé par (N – 1) ; c'est ce qui est fait dans la plupart des exemples subséquents.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Écart type (S). Comme la variance est calculée à partir des notes de déviation élevées au carré, il faut en extraire la racine carrée pour obtenir un indice exprimé dans les mêmes unités que les notes de déviation ou les notes brutes. La racine carrée de 13,63 est de 3,69, soit l'écart type des notes de la classe. L'écart type permet de connaître l'étalement ou la dispersion d'une série d'observations. Évidemment, plus l'écart type est faible, plus les notes se distribuent étroitement autour de la moyenne, et inversement ; c'est un indice de dispersion ou d'homogénéité des données. Un indice de dispersion peut aussi servir à apprécier dans quelle mesure la moyenne est un résumé exact et utile de l'ensemble des observations. En effet, ce résumé fourni sera d'autant représentatif que la dispersion des observations est faible, c'est-à-dire que les observations sont réparties dans l'entourage immédiat de la moyenne.

Coefficient de variation (CV). L'écart type est un indice qui doit s'interpréter par rapport aux mêmes unités que les notes brutes. Comment alors comparer des écarts types calculés sur des ensembles de données qui ne sont pas exprimées dans les mêmes unités ? Par exemple, l'écart type de 3,69 de l'examen final (sur 30) traduit-il une dispersion plus grande ou plus petite que l'écart type de 3,00 de l'examen intra (sur 20) ? Il faut recourir à une mesure de dispersion relative, qui est le coefficient de variation ; il s'obtient en divisant l'écart type par la moyenne, puis en multipliant le résultat par 100 afin d'exprimer l'indice en pourcentage. La formule est la suivante :

Coefficient de variation $(\mathrm{CV}) = \dfrac{S_x}{M_x} \times 100$
où S_x : écart type des X_i M_x : moyenne des X_i

Le coefficient de variation, en étant indépendant de l'unité de mesure des données, permet de comparer la dispersion d'ensembles de données qui ne sont pas exprimées dans la même unité ou dont les moyennes sont très différentes (Baillargeon, 1989). En supposant que la moyenne de l'examen intra soit de 14,1, on obtient un coefficient de variation de 21 % (soit l'écart type 3,00 divisé par la moyenne 14,1 et multiplié par 100) ; cela dépasse légèrement le

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 7.4
COEFFICIENT DE VARIATION ET DIFFÉRENCE D'UNITÉS

	Examen intra	**Examen final**
Unités de mesure	Échelle de 0 à 20	Échelle de 0 à 30
Moyenne (M)	14,10	21,33
Écart type (S)	3,00	3,69
Coefficient de variation (CV)	$\frac{3,00}{14,1} \times 100 = 21\,\%$	$\frac{3,69}{21,33} \times 100 = 17\,\%$

coefficient de variation de 17 % de l'examen final (soit l'écart type 3,69 divisé par la moyenne 21,33, puis multiplié par 100). Les calculs sont présentés au tableau 7.4.

Signalons que le coefficient de variation ne doit être utilisé que pour des données dont le niveau de mesure appartient à une échelle de rapport (Baillargeon et Martin, 1994), ce qui implique des données dont l'origine absolue est zéro (revoir section « Échelles de mesure »). Les moyennes provenant de distributions différentes peuvent alors servir de dénominateur et l'on peut ainsi les rendre comparables parce qu'elles ont toutes une origine semblable, à savoir zéro. En revanche, si les origines diffèrent et que les données sont du niveau « échelle d'intervalle », il faut corriger la différence d'origine en modifiant les moyennes par des constantes appropriées avant de procéder au calcul des coefficients de variation. Le plus simple est de ramener les moyennes comme si l'origine réelle était zéro.

Voyons l'exemple donné au tableau 7.5. Supposons que l'on veuille comparer la dispersion d'un examen de connaissance du français avec celle d'un test d'aptitudes. Les résultats à l'examen de français sont en pourcentage, avec une moyenne de 75 % et un écart type de 10 %. La distribution des résultats se situe systématiquement entre 45 % et 95 %, même si, en théorie, elle aurait pu occuper toute l'échelle de 100 points. Les résultats au test d'aptitudes sont en QI (quotient intellectuel) et peuvent varier pour ce test entre 40 et 160, avec une moyenne de 100 et un écart type de 20. Le calcul des coefficients de variation sont respectivement de 13 % pour l'examen (soit l'écart type 10 divisé par la moyenne 75, multiplié par 100) et de 20 % pour le test (soit l'écart type 20 divisé par la moyenne 100,

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 7.5
CORRECTION DU COEFFICIENT DE VARIATION LORSQUE LES ORIGINES DES DISTRIBUTIONS NE SONT PAS SEMBLABLES

	Examen de français	**Test d'aptitudes**
Unités de mesure	Pourcentage	QI
Moyenne (M)	75 %	100
Écart type (S)	10 %	20
Résultat minimum	45 %	40
Résultat maximum	95 %	160
A) Sans correction		
Coefficient de variation (CV)	$\frac{10}{75} \times 100 = 13\,\%$	$\frac{20}{100} \times 100 = 20\,\%$
B) Avec correction		
Coefficient de variation (CV)	$\frac{10}{75-45} \times 100 = 33\,\%$	$\frac{20}{100-40} \times 100 = 33\,\%$

multiplié par 100). On serait porté à croire que la dispersion à l'examen de français est plus faible que celle du test d'aptitudes, alors qu'en fait elle est identique.

Pour s'en convaincre, il s'agit de reprendre les calculs en plaçant l'origine de chacune des distributions des résultats au même niveau, en l'occurrence à zéro. Comme la distribution des résultats à l'examen débute à 45, il faut retrancher cette distance de la moyenne actuelle de 75, pour donner une moyenne corrigée de 30. Cela entraîne un coefficient corrigé de variation de 33 % pour l'examen de français (soit l'écart type 10 divisé par la moyenne corrigée 30 multiplié par 100). Quant au test d'aptitudes, la moyenne actuelle de 100 doit être décalée de 40 points vers la gauche pour donner un coefficient corrigé de variation également de 33 % (soit l'écart type 20 divisé par la moyenne corrigée 60 et multiplié par 100).

NOTE STANDARD (Z)

Est-ce qu'un score de 110 est toujours meilleur qu'un score de 85 % ? Bien sûr que non. Il faut d'abord savoir s'il s'agit de la même unité de mesure. Est-ce qu'un score de 85 % dans un groupe est toujours meilleur qu'un score de 65 % dans un autre groupe ? Non encore une fois : cela dépend de la moyenne de chacun des groupes. Est-ce qu'au moins deux scores de 85 % sont égaux, si les moyennes des deux distributions sont égales ? Non, il faut connaître les écarts types.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Lorsque les observations proviennent du même ensemble, il est aisé de les comparer entre elles, qu'elles soient sous forme de notes brutes ou de scores de déviation ; la comparaison directe est possible parce que chaque observation fait partie de la même distribution. Il en va tout autrement si les observations n'ont pas une unité de mesure, une moyenne et un écart type identiques. La note standard (Z) permet de contourner ces difficultés ; elle s'obtient en divisant la note de déviation par l'écart type, comme l'exprime la formule suivante :

Note standard $\left(Z_i\right) = \dfrac{X_i - M_x}{S_x}$

où Z_i : note standard pour X = i

X_i : chaque observation (personne, objet ou événement)

M_x : moyenne des X_i

S_x : écart type des X_i

Les exemples compilés au tableau 7.6 illustrent comment la note standard nivelle les différences de moyenne, d'écart type et d'unité de mesure lors de la comparaison de données ne provenant pas d'un même ensemble. Dans l'exemple A, où seules les **moyennes** diffèrent, on peut voir que la note de déviation aurait suffi à établir une comparaison équitable. Dans les deux autres exemples, cependant, on constate que la note de déviation n'y suffit pas. Si les écarts types ou l'unité de mesure ne sont pas identiques, il faut absolument recourir à la note standard pour rectifier les distorsions dans les comparaisons. La note standard, comme son nom l'indique, appartient à une échelle standardisée, qui est toujours la même. Peu importe la nature des notes brutes, leur conversion en notes standard va centrer leur distribution autour d'une moyenne égale à 0 et ramener l'écart type à 1, de sorte qu'une unité de note standard vaut un écart type. Une note standard est simplement la note brute exprimée en nombre d'écarts types au-dessus ou en dessous de la moyenne (voir tableau 7.2, dernière colonne).

Revenons au tableau 7.6. L'exemple B montre comment la note standard permet de contourner une différence **d'écart type**. Dans cet exemple, la note de 10 points au-dessus de la moyenne vaut 1,00 en note standard lorsque l'écart type des notes brutes est de 10 (examen de français), mais vaut 2,00 en note standard lorsque l'écart type des notes brutes est de 5 (examen d'histoire). Dans l'exemple C, on

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 7.6
NOTE STANDARD ET DIFFÉRENCE DE MOYENNE, D'ÉCART TYPE OU D'UNITÉ DE MESURE

A) Différence de moyenne

	Examen de français	Examen d'arithmétique
Unités de mesure	Pourcentage	Pourcentage
Moyenne (M)	75 %	55 %
Écart type (S)	10 %	10 %
Note brute (X_i)	85 %	65 %
Note de déviation ($X_i - M_x$)	10 %	10 %
Note standard (Z_i) = ($X_i - M_x$)/S_x	1,00 %	1,00 %

B) Différence d'écart type

	Examen de français	Examen d'histoire
Unités de mesure	Pourcentage	Pourcentage
Moyenne (M)	75 %	75 %
Écart type (S)	10 %	5 %
Note brute (X_i)	85 %	85 %
Note de déviation ($X_i - M_x$)	10 %	10 %
Note standard (Z_i) = ($X_i - M_x$)/S_x	1,00 %	2,00 %

C) Différence d'unité de mesure

	Examen de français	Test d'aptitudes
Unités de mesure	Pourcentage	QI
Moyenne (M)	75 %	100
Écart type (S)	10 %	20
Note brute (X_i)	85 %	110
Note de déviation ($X_i - M_x$)	10 %	10
Note standard (Z_i) = ($X_i - M_x$)/S_x	1,00 %	0,50

compare deux résultats qui ne sont pas exprimés dans la même **unité de mesure**. Grâce à la note standard, on réalise qu'une note brute de 85 % à l'examen de français, qui vaut 1,00 en note standard, est plus forte qu'une note de 110 de QI à un test d'aptitudes qui n'équivaut qu'à 0,50 en note standard.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Malgré les avantages de la note standard, cette transformation donne lieu à des résultats qui sont **relatifs aux groupes de référence** respectifs et, par conséquent, les comparaisons d'une note à l'autre ne sont possibles que si elles proviennent du même groupe ou de groupes équivalents de personnes. Reprenons l'exemple B. La note standard de 1,00 à l'examen de français signifie que la personne se situe à 1,00 écart type au-dessus de la moyenne de cette classe, et la note standard de 2,00 à l'examen d'histoire vaut 2,00 écarts types au-dessus de cette classe en histoire. Que ces classes de français et d'histoire soient constituées d'étudiants forts ou faibles, cela ne modifie en rien l'interprétation de la note standard par rapport à chaque classe : la note standard est la position d'un résultat en fonction de la performance du groupe. Mais, si l'on veut comparer deux notes standard provenant de groupes différents, l'équivalence des groupes devient primordiale. Par exemple, si la classe de français regroupe des étudiants très forts dans cette matière alors que la classe d'histoire est formée d'étudiants plutôt faibles en histoire, il est plus difficile pour un étudiant d'obtenir une note au-dessus de la moyenne en français comparativement en histoire.

DISTRIBUTION NORMALE

Il arrive parfois que l'histogramme des observations d'une variable quantitative suive une répartition semblable à une distribution en forme de cloche, comme celle rapportée à la figure 7.3. Cet histogramme a été compilé à partir de la distribution de fréquences des résultats au test psychométrique Otis mesurant l'aptitude mentale générale pour 253 travailleurs dans une usine. On constate que les résultats de la majorité des individus se regroupent autour de la moyenne (M = 103) et que les résultats au-dessus et au-dessous sont de moins en moins nombreux à mesure que l'on s'éloigne de la moyenne. La forme de cette répartition n'est pas sans rappeler la distribution normale (voir figure 7.4).

Propriétés mathématiques de la courbe normale. La distribution normale, appelée aussi courbe normale ou loi normale, est une distribution de fréquences qui, par définition, comporte d'importantes propriétés mathématiques (Baillargeon, 1989). Compte tenu de la nature pratique de notre démarche, seulement quelques-unes d'entre elles seront présentées. Premièrement, avec sa forme de cloche parfaitement **symétrique**, les observations se répartissent de part et d'autre

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 7.3
RÉSULTATS DE 253 TRAVAILLEURS EN USINE AU TEST OTIS

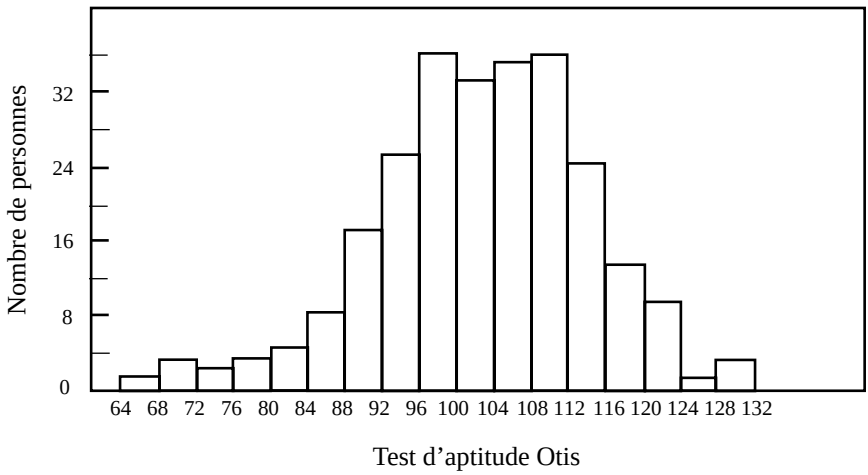
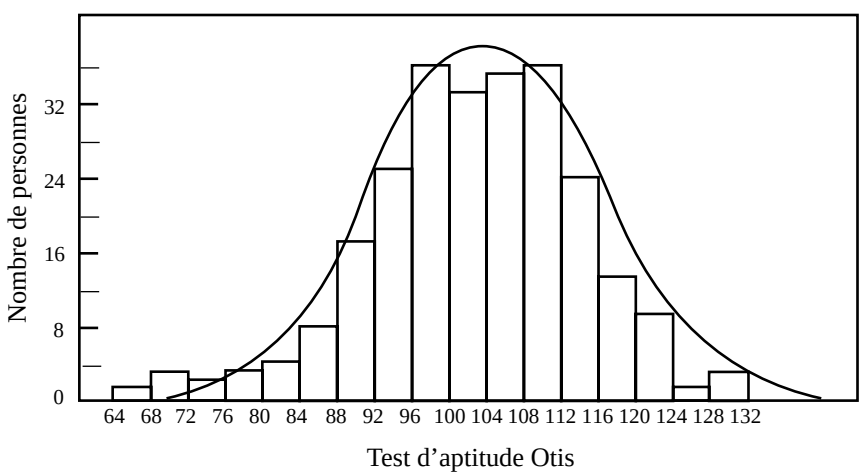


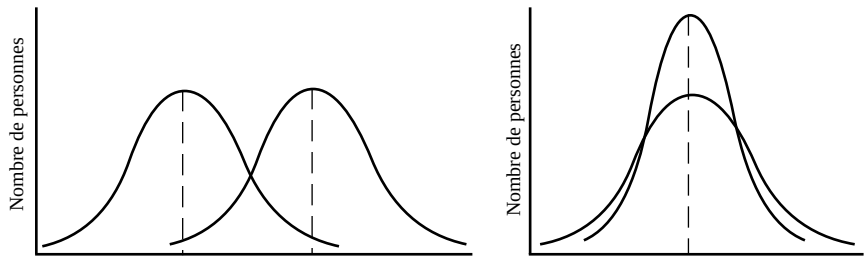
Figure 7.4
RÉSULTATS DE 253 TRAVAILLEURS EN USINE AU TEST OTIS EN FONCTION DE LA DISTRIBUTION NORMALE



© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 7.5
DISTRIBUTIONS NORMALES DÉFINIES PAR LEUR MOYENNE ET LEUR ÉCART TYPE



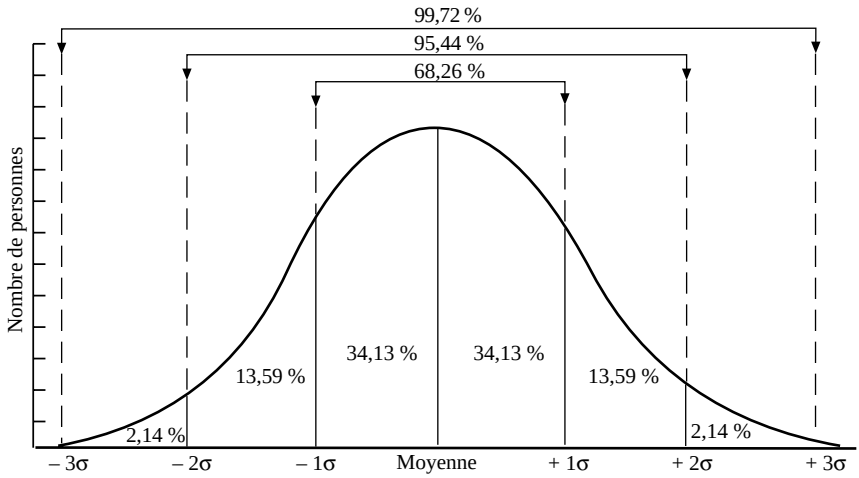
de la moyenne qui figure exactement au centre (voir figure 7.4). Deuxièmement, elle est **définie par sa moyenne (μ) et son écart type(σ)**[3]. Deux distributions normales peuvent avoir le même écart type, donc être de même forme, mais avoir des moyennes différentes (voir figure 7.5, partie gauche) ; réciproquement, elles peuvent avoir la même moyenne, mais des écarts types différents. Plus l'écart type est élevé, plus la distribution est aplatie (voir figure 7.5, partie droite).

Troisièmement, et c'est là une des propriétés les plus importantes, le **pourcentage d'observations est connu** pour tout intervalle de la variable (voir figure 7.6). Ainsi, par définition, 34,13 % des observations se trouvent entre la moyenne et un écart type, ce qui fait 68,26 % des observations entre un écart type en dessous de la moyenne et un écart type au-dessus de la moyenne. De la même manière, il y a 95,44 % des observations entre deux écarts types en dessous de la moyenne et deux écarts types au-dessus de la moyenne. Enfin, 99,72 % des observations sont comprises entre trois écarts types au-dessous de la moyenne et trois écarts types au-dessus de la moyenne. En apparence insignifiantes, les décimales des pourcentages d'observations doivent être prises en considération, surtout si la taille de la population est grande. Par exemple, même si le pourcentage d'observations au-delà des trois écarts types n'est que de 0,28 % de la distribution (soit 100 % – 99,72 %), ce n'est pas nécessairement négligeable. Dans le cas d'une population d'une agglomération comme Trois-Rivières (120 000 citoyens) ou Montréal (3 000 000 de citoyens),

3. La moyenne et l'écart type sont représentés par les symboles μ et σ parce qu'il s'agit de paramètres d'une distribution hypothétique.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 7.6
RÉPARTITION DES OBSERVATIONS DANS UNE DISTRIBUTION NORMALE



il y a respectivement 336 ou 8400 personnes qui se retrouvent dans cette zone. Étant donné que la courbe normale est une asymptote, elle ne rejoint, en théorie, jamais l'axe horizontal.

Il est souvent plus pratique d'avoir des pourcentages d'observations **sans décimale**, comme lorsqu'on veut savoir entre quelles valeurs se retrouve 95 % de la population. Dans ce cas, il faut retrancher les quelque 0,44 % d'observations (95,44 % – 0,44 % = 95 %) en se rapprochant légèrement vers la moyenne. Ainsi, pour contenir 95 % des observations, on retient 1,96 écart type en bas de la moyenne et 1,96 écart type en haut de la moyenne. Pour 99 % des observations, il faut prendre 2,58 écarts types en bas et 2,58 écarts types en haut[4].

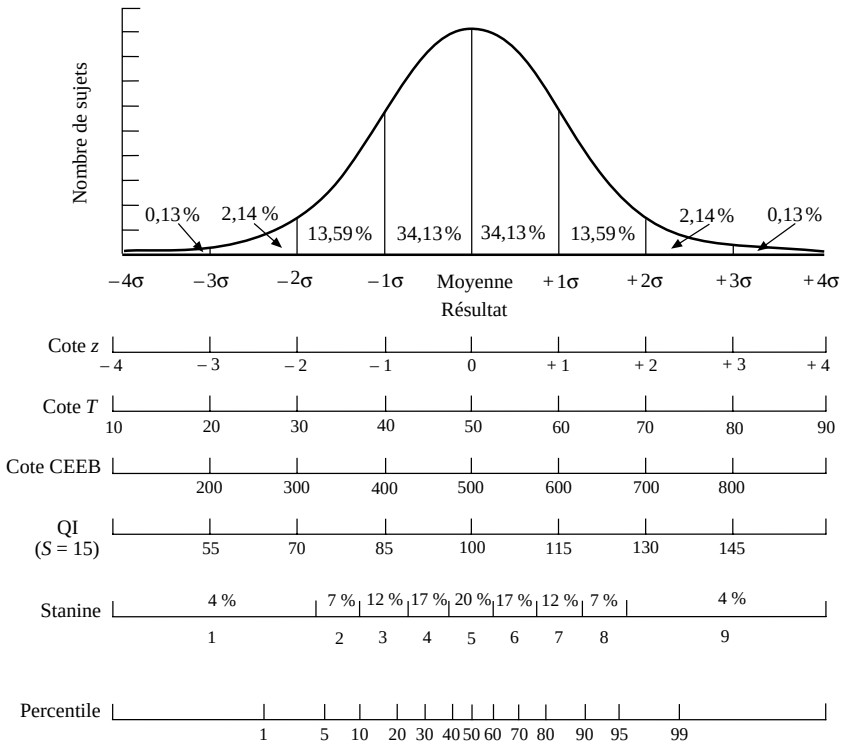
Examinons la répartition des observations dans une distribution normale en fonction de **divers types de notes** (voir figure 7.7) ; il est aisé de constater que les notes standard (Z) sont des valeurs dont

4. Pour connaître la proportion d'observations correspondant à toute autre valeur de la variable, il faut s'en remettre à la table de la loi normale dite centrée réduite. Appelée aussi distribution de Z, la loi normale centrée réduite est une distribution dont les valeurs sont établies en notes standard ; cette table est reproduite à l'appendice A.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 7.7
RELATION ENTRE LES DIFFÉRENTS TYPES DE NOTES EN FONCTION DE LA DISTRIBUTION NORMALE



l'unité correspond à un écart type et dont la moyenne est zéro. Regardons aussi les percentiles, dont les notes expriment le pourcentage de cas ayant une valeur plus faible qu'une note brute donnée. Si 15 % des personnes obtiennent à un examen une note brute inférieure à 60, à cette note brute de 60 correspond un percentile de 15. Par rapport à la distribution normale, on remarque que le rang percentile 50 correspond au milieu de la courbe, indiquant que 50 % de l'échantillon se trouve au-dessus de cette note et que 50 % se trouve en dessous. Ou bien, à une note Z de +1 correspond un rang percentile d'environ 84 (soit 50 % + 34,13 % = 84,13 %).

Applications pratiques de la courbe normale. La distribution normale parfaite telle qu'elle est décrite ci-dessus est un modèle théorique aux applications fort nombreuses. En **statistique**, elle est

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

fondamentale puisqu'elle sert à calculer les probabilités d'innombrables phénomènes aléatoires. En effet, Quetelet, un astronome et mathématicien belge, avait noté que la nature a tendance à errer de part et d'autre du centre de la distribution, suivant ce qui s'appelait alors la « loi normale de l'erreur » (Guion, 1998). La fidélité et l'erreur type de mesure (voir chapitre 4) en sont des applications directes.

En **gestion des ressources humaines**, la distribution normale a également une grande importance, car elle représente la répartition des observations pour plusieurs phénomènes et caractéristiques individuelles. Par exemple, la majorité des tests psychométriques mesurant les aptitudes, les traits de personnalité, les intérêts ou autres caractéristiques individuelles donnent lieu à ce genre de distribution. Les premiers psychométriciens ont adopté la distribution normale pour calibrer leurs tests parce qu'ils croyaient que les aptitudes et autres caractéristiques fondamentales se distribuaient selon la même loi naturelle qui régit plusieurs caractéristiques physiques ou biologiques dans une population. En effet, des caractéristiques telles que la taille ou le poids des personnes appartenant à une population donnée ont tendance à se distribuer normalement. Selon Cronbach (1970), la dimension aléatoire propre à ce genre de distribution proviendrait probablement du facteur chance qui intervient lors de la combinaison des chromosomes, lesquels sont responsables d'une grande partie de la variation d'une personne à l'autre par rapport à un type de caractéristiques.

La courbe normale est devenue et demeure un moyen pratique d'étalonner la plupart des tests psychométriques mesurant une caractéristique humaine (Guion, 1998 ; Murphy et Davidshofer, 1988). Par conséquent, lorsqu'un test psychométrique est appliqué 1) à des personnes représentatives de la population, c'est-à-dire en nombre suffisamment grand et sélectionnées au hasard et 2) que le test n'est ni trop facile ni trop difficile, la distribution des résultats obtenus s'apparente à la courbe normale (Maier, 1970). La figure 7.4 illustre ce phénomène, même si la distribution n'est pas parfaitement normale. Il faut rappeler dans cet exemple que la taille de l'échantillon est de 253 personnes et qu'il s'agit de travailleurs qui ne sont pas complètement choisis au hasard parmi l'ensemble de la population. Néanmoins, la distribution des résultats a nettement l'apparence d'une distribution normale. Dans le cas d'échantillons plus petits, il

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

ne faudrait pas s'attendre à ce que les distributions prennent l'allure de la distribution normale théorique ; seules quelques tendances seraient prévisibles (revoir la figure 7.2 en guise d'exemple).

DÉVIATION PAR RAPPORT À LA DISTRIBUTION NORMALE

Si un caractère ou un comportement est mesuré chez un groupe de personnes et que la distribution n'est pas normale, il y a au moins cinq **explications** possibles à ce phénomène.

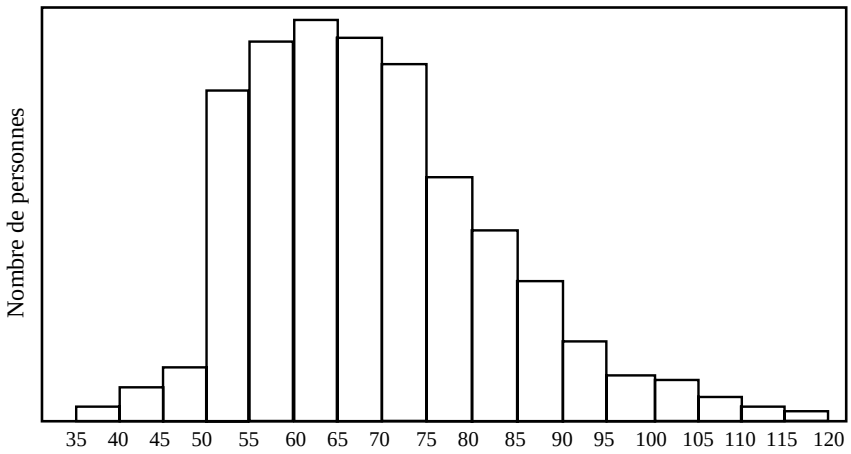
Erreur d'échantillonnage. Il se peut que cela soit simplement dû au hasard, comme c'est souvent le cas avec un échantillon trop petit. Cette erreur due à l'échantillonnage est vue dans ce chapitre à la section « Coefficients de corrélation et de détermination », et dans le chapitre 3, aux sections « Considérer les faiblesses méthodologiques », « Double validation » et « Méta-analyse, généralisation de la validité et autres méthodes de validation ».

Groupe biaisé. Il est possible que les personnes ne soient pas tirées au hasard d'une population donnée et que l'on soit conséquemment en présence d'un groupe biaisé. Par exemple, supposons qu'un examen de dactylographie soit administré à un groupe de secrétaires à l'emploi d'une organisation et que l'on obtienne la distribution apparaissant à la figure 7.8. Cette distribution, asymétrique ou biaisée vers la droite, autorise à penser que ce groupe de secrétaires n'est pas représentatif de la population en général. En effet, leurs résultats à l'examen de dactylographie ne se répartissent pas également de part et d'autre de la majorité des notes situées aux environs de 50 à 75 mots à la minute. Alors que le nombre de secrétaires diminue progressivement à mesure que l'on s'éloigne de la majorité vers la droite, il chute lorsqu'on se dirige vers la gauche à partir de 50 mots à la minute. Dans ce cas fictif, l'explication réside dans le fait que la plupart des candidates à ce poste qui n'ont pas réussi au moins 50 mots à la minute lors d'un examen semblable à l'embauche n'avaient pas été sélectionnées.

Populations différentes. Il arrive que deux populations différentes soient réunies dans une même distribution ; on reconnaît souvent cette distribution à sa forme bimodale, c'est-à-dire une distribution ayant deux sommets (Maier, 1970). En guise d'exemple, on peut imaginer ce qui se passerait si un examen de français était appliqué à un échantillon de candidats composés de francophones et d'allophones.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 7.8
RÉSULTATS À UN EXAMEN DE DACTYLOGRAPHIE D'UN GROUPE DE SECRÉTAIRES AYANT DÉJÀ ÉTÉ SÉLECTIONNÉES

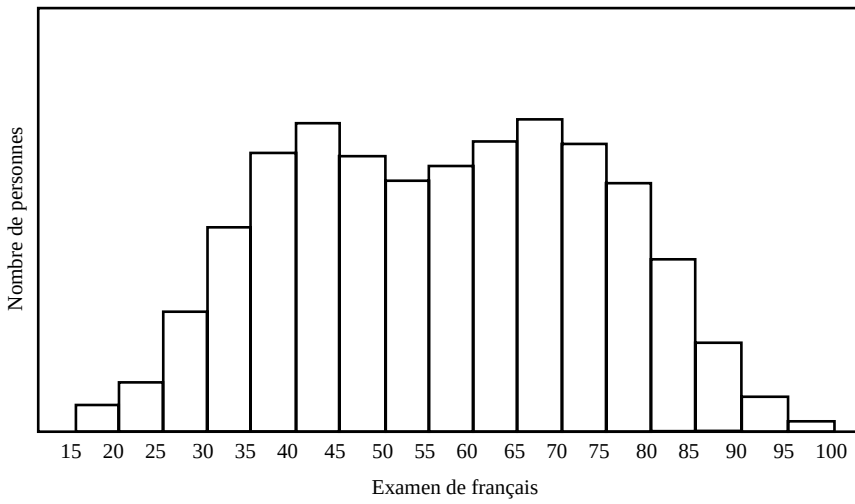


La distribution obtenue aurait vraisemblablement deux sommets, l'un pour les allophones et l'autre pour les francophones, correspondant grossièrement à la moyenne de chaque groupe (voir figure 7.9). Une distribution bimodale signale la présence de deux populations ; elle peut aussi révéler une séparation dans une population, ce qui revient au même. Maier (1970) donne l'exemple d'un échantillon de travailleurs parmi lesquels ceux qui avaient un rendement moyen se sont démotivés pour une raison quelconque. La distribution de l'ensemble de l'échantillon laisserait voir deux sommets, l'un pour les employés à haut rendement et l'autre pour les employés à bas rendement auxquels se sont ajoutés ceux qui ont perdu leur motivation.

Facteurs externes. Des facteurs externes peuvent influencer les résultats et ainsi modifier la distribution normale. Par exemple, lors d'une intervention dans une usine de fabrication de papier, la compilation des évaluations du rendement des employés de production a donné une distribution semblable à celle de la figure 7.10. Presque tous les employés recevaient de leur superviseur une évaluation de trois (« Satisfait aux exigences ») sur une échelle en quatre points. Une des hypothèses d'explication qui a été examinée est que les superviseurs donnaient la même évaluation à presque tout le monde pour

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 7.9
RÉSULTATS À UN EXAMEN DE FRANÇAIS PASSÉ PAR DES FRANCOPHONES ET DES ALLOPHONES

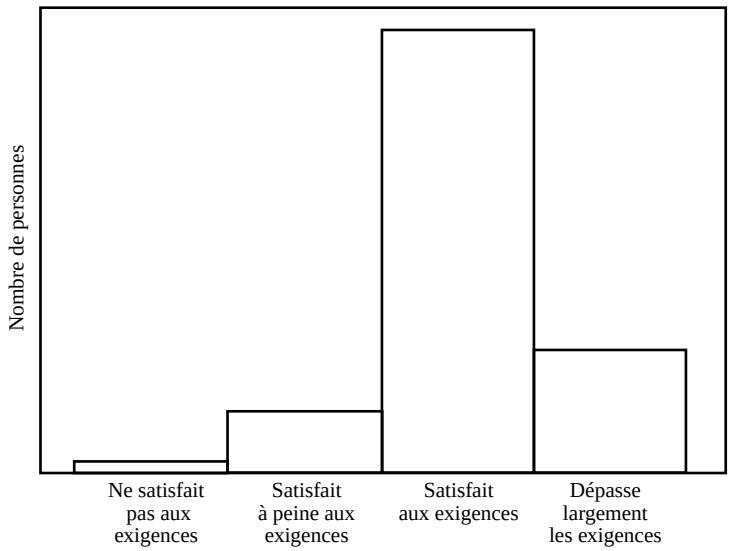


s'éviter des ennuis. Fermer les yeux sur les cas de rendement insuffisant pouvait être un moyen de ne pas susciter la colère et l'agressivité possibles de la part des employés. Réciproquement, ne pas reconnaître les employés les plus méritants pouvait dispenser les superviseurs d'avoir à justifier auprès de la majorité des employés pourquoi ils ne font que « satisfaire aux exigences ». De telles considérations, d'ordre « politique », constituent souvent des facteurs extérieurs qui modifient la forme de la distribution des évaluations du rendement.

Nature de la caractéristique ou du phénomène. Une multitude de caractéristiques ou de phénomènes ne se distribuent pas selon la courbe normale. Par exemple, l'une des croyances erronées est que le rendement des employés se distribue de façon normale dans une entreprise, alors qu'en réalité la véritable distribution du rendement est presque toujours inconnue (Gosselin et Murphy, 1994). En fait, la gestion des ressources humaines a précisément comme préoccupation que cette distribution ne soit pas normale. En effet, par ses pratiques de recrutement, de formation ou de gestion du rendement, l'organisation tente de modeler le rendement de ses employés à la hausse. Bref, interpréter une distribution exige non seulement des compé-

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 7.10
RÉSULTATS À L'ÉVALUATION DU RENDEMENT POUR LES EMPLOYÉS D'UNE USINE



tences statistiques, mais aussi des connaissances sur la théorie sous-jacente au concept mesuré ainsi que sur le contexte dans lequel les mesures ont été effectuées.

Revenons à l'exemple précédent (figure 7.10). Pour bien analyser les causes qui donnent cette forme particulière à la distribution des résultats à l'évaluation du rendement, il aurait absolument fallu examiner les politiques et les objectifs de l'organisation en matière d'évaluation du rendement. Peut-être aurions-nous découvert quelques facteurs extérieurs qui auraient infirmé l'hypothèse d'un laisser-aller de la part des superviseurs. Par exemple, peut-être que l'organisation vise à encourager l'esprit d'équipe et d'entraide entre les employés, plutôt que la compétition individuelle. Conséquemment, le système d'évaluation du rendement ne sert qu'à détecter les cas extrêmes nécessitant une intervention qui va au-delà de la gestion quotidienne. L'échelle servant à évaluer les employés n'avait donc pas été conçue pour discriminer plus finement la majorité des employés qui se retrouvent de part et d'autre de la moyenne. La catégorie « Dépasse largement les exigences » renvoie à un rendement exceptionnel alors

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

que les catégories « Satisfait à peine aux exigences » et « Ne satisfait pas aux exigences » correspondent à un rendement d'une faiblesse effectivement peu fréquente.

RELATION ENTRE DEUX VARIABLES

Jusqu'ici, les variables ont été considérées une à une, sans les mettre en relation entre elles. Or, en gestion des ressources humaines, il est souvent primordial de connaître la relation entre deux variables. Par exemple, dans quelle mesure les résultats scolaires sont-ils un bon indice du rendement de la personne au travail ? Est-ce essentiel d'avoir travaillé cinq ans dans le secteur pour être capable d'effectuer les tâches reliées à ce poste de représentant ? Est-ce que l'on peut se fier au dynamisme démontré par un candidat lors d'un entrevue de sélection pour prédire son comportement une fois embauché ? Est-ce que l'évaluation d'un interviewer est corroborée par celle des autres interviewers ? Est-ce que la perception qu'a un superviseur concernant un employé correspond à celle de la clientèle à son égard ? Les outils statistiques les plus utilisés pour analyser la relation entre deux variables (de niveau « échelle d'intervalle » ou « échelle de rapport ») sont le diagramme de dispersion et le coefficient de corrélation.

DIAGRAMME DE DISPERSION

Commençons par un exemple. Le directeur du programme de sciences comptables est celui qui doit aider les finissants à décider s'ils vont se présenter aux examens de l'Ordre des comptables agréés afin d'obtenir ce titre professionnel. Cependant, tout étudiant désireux de se présenter aux examens et de devenir comptable agréé (c.a.) doit considérer qu'il aura obligatoirement une période de formation intensive additionnelle d'une durée de plus de trois mois au cours de l'été. Même si cette formation est très exigeante, en plus d'empêcher les étudiants de toucher un revenu en occupant un emploi, elle ne garantit pas le succès : le taux d'échec national oscille, bon an mal an, autour de 50 %. L'étudiant doit donc peser le pour et le contre avant de commencer un tel processus. Si seulement l'étudiant pouvait évaluer ses chances de réussite, cela faciliterait son processus décisionnel. Devant cette problématique, le directeur du module a la bonne idée d'examiner la relation entre les résultats au baccalauréat et la

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

note obtenue aux examens de l'Ordre des comptables agréés pour ses étudiants des années antérieures. En effet, il a souvent observé que les meilleurs étudiants au baccalauréat avaient tendance à mieux réussir les examens de l'Ordre.

Pour les diplômés de 1980 qui se sont présentés aux examens, il a recueilli pour chacun d'eux leur moyenne cumulative au baccalauréat en sciences comptables et leur résultat aux examens (voir tableau 7.7, de la première à la troisième colonne). L'analyse préliminaire des **notes brutes** révèle que la moyenne cumulative pour ces 46 diplômés varie de 2,0 à 3,9 et que le résultat aux examens de l'Ordre varie de 150 à 340. Au-delà de ces constatations, il n'est pas facile de détecter s'il y a une relation entre ces deux variables. La transformation des données brutes en **notes standard** laisse entrevoir l'existence possible d'une relation entre la moyenne universitaire et les résultats aux examens de l'Ordre (voir tableau 7.7, quatrième et cinquième colonnes). Comme nous l'avons mentionné plus tôt, la note standard (Z) facilite grandement la comparaison de données provenant de deux distributions différentes. Ainsi, il est aisé de constater pour le diplômé n° 1 que sa moyenne universitaire de –1,58 prédit assez bien son résultat de –2,06 aux examens. Il en est de même pour plusieurs diplômés (n° 2, n° 3, n° 4, n° 7, etc.). Pour d'autres, en revanche, la moyenne universitaire est moins révélatrice (n° 12, n° 16, n° 30, n° 37, etc.). Même si la note standard facilite la comparaison des données prises deux à deux, il est difficile d'avoir une vue d'ensemble lorsqu'il y a plusieurs couples de données. Aussi est-il préférable de compiler un diagramme de dispersion.

Description. Un diagramme de dispersion est simplement un tableau à deux axes, où chaque couple d'observations (X_i ; Y_i) est représenté à l'aide d'un point situé à l'intersection des coordonnées correspondant à la valeur de X_i et de Y_i. Dans cet exemple (voir figure 7.11), la moyenne cumulative universitaire constitue la variable indépendante. Par convention, la variable indépendante est notée X et elle apparaît sur l'axe horizontal (ou en abscisse). Les résultats aux examens est la variable dépendante ; elle est notée Y et elle est située sur l'axe vertical (ou en ordonnée). Ainsi, le diplômé n° 1 (voir tableau 7.7) est représenté dans le diagramme de dispersion (voir figure 7.11) par le point qui se situe à la fois vis-à-vis de la valeur 2,0 sur l'axe horizontal et vis-à-vis de 150 sur l'axe vertical : c'est le point que l'on voit dans le coin inférieur gauche. Le finissant n° 2 est situé

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau 7.7

MOYENNE CUMULATIVE UNIVERSITAIRE ET RÉSULTAT AUX EXAMENS DE L'ORDRE DES COMPTABLES AGRÉÉS POUR 46 DIPLÔMÉS EN 1980

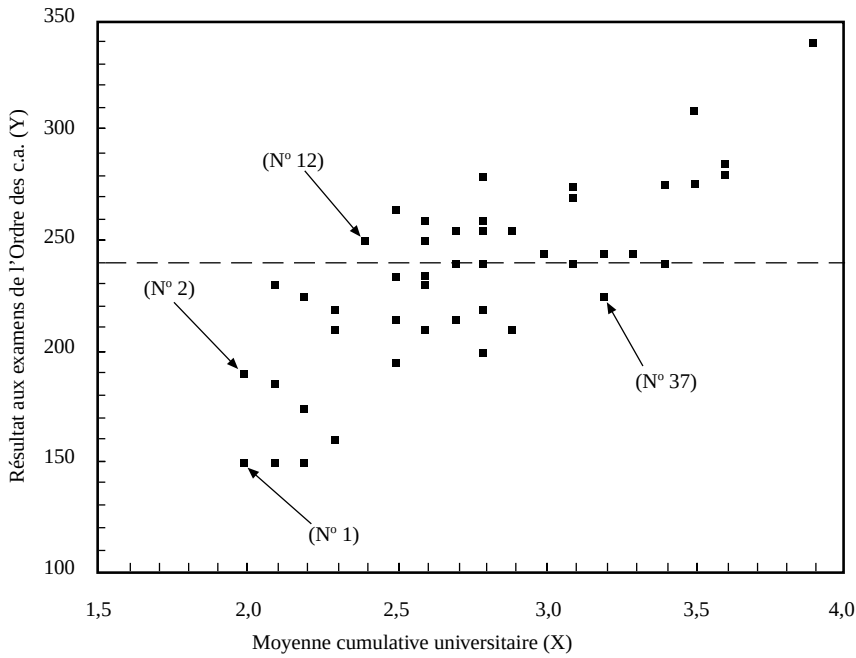
Diplômé	Moyenne universitaire (X_i)	Résultats aux examens de l'Ordre (Y_i)	Z_{x_i} (Moy. univ.)	Z_{y_i} (Examens)	Produit de Z_{x_i} Z_{y_i}
1	2,0	150	–1,58	–2,06	3,26
2	2,0	190	–1,58	–1,08	1,71
3	2,1	150	–1,38	–2,06	2,83
4	2,1	185	–1,38	–1,20	1,65
5	2,1	230	–1,38	–0,10	1,14
6	2,2	150	–1,17	–2,06	–2,40
7	2,2	175	–1,17	–1,45	1,69
8	2,2	225	–1,17	–0,22	0,26
9	2,3	160	–0,96	–1,82	1,74
10	2,3	210	–0,96	–0,59	0,57
11	2,3	220	–0,96	–0,35	0,33
12	2,4	250	–0,75	0,39	–0,29
13	2,5	195	–0,54	–0,96	0,52
14	2,5	215	–0,54	–0,47	0,25
15	2,5	235	–0,54	0,02	–0,01
16	2,5	265	–0,54	0,76	–0,41
17	2,6	210	–0,33	–0,59	0,20
18	2,6	230	–0,33	–0,10	0,03
19	2,6	235	–0,33	0,02	–0,01
20	2,6	250	–0,33	0,39	–0,13
21	2,6	260	–0,33	0,63	–0,21
22	2,7	215	–0,12	–0,47	0,06
23	2,7	240	–0,12	0,14	–0,02
24	2,7	255	–0,12	0,51	–0,06
25	2,8	200	0,08	–0,84	–0,07
26	2,8	220	0,08	–0,35	–0,03
27	2,8	240	0,08	0,14	0,01
28	2,8	255	0,08	0,51	0,04
29	2,8	260	0,08	0,63	0,05
30	2,8	280	0,08	1,12	0,09
31	2,9	210	0,29	–0,59	–0,17
32	2,9	255	0,29	0,51	0,15
33	3,0	245	0,50	0,27	0,13
34	3,1	240	0,71	0,14	0,10
35	3,1	270	0,71	0,88	0,62
36	3,1	275	0,71	1,00	0,71
37	3,2	225	0,92	–0,22	–0,20
38	3,2	245	0,92	0,27	0,24
39	3,3	245	1,13	0,27	0,30
40	3,4	240	1,33	0,14	0,19
41	3,4	275	1,33	1,00	1,33
42	3,5	275	1,54	1,00	1,54
43	3,5	310	1,54	1,86	2,86
44	3,6	280	1,75	1,12	1,97
45	3,6	285	1,75	1,25	2,18
46	3,9	340	2,38	2,59	6,16
Moyenne (M) =	2,76	234,13	–0,01	0,00	0,76
Écart type (S) =	0,48	40,83	1,00	1,00	

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 7.11
DIAGRAMME DE DISPERSION ENTRE LA MOYENNE CUMULATIVE UNIVERSITAIRE ET LE RÉSULTAT AUX EXAMENS DE L'ORDRE DES COMPTABLES AGRÉÉS POUR 46 CANDIDATS



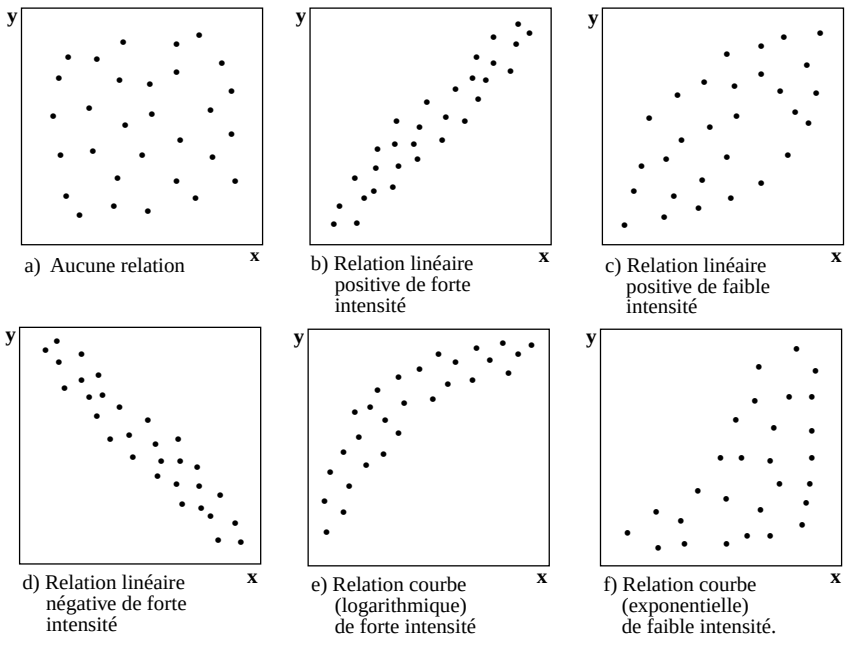
à l'intersection des valeurs 2,0 et 190. Et ainsi de suite pour chacun des 46 finissants. Il y a donc autant de points que de personnes dans l'échantillon et chaque point représente simultanément la moyenne cumulative universitaire, soit la valeur X_i, et le résultat aux examens de l'Ordre des comptables agréés, soit la valeur Y_i. Comme nous l'avons indiqué plus haut, cette figure montre que la moyenne cumulative varie de 2,0 à 3,9 et que le résultat aux examens va de 150 à 340. Mais que peut-on y voir de plus ? Peut-on déceler une relation entre la moyenne cumulative universitaire et les examens de l'Ordre[5] ?

5. Le diagramme de dispersion, compilé à partir des notes standard au lieu des notes brutes, aurait donné le même nuage de points, seule l'échelle des axes aurait été changée.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 7.12
DIAGRAMMES DE DISPERSION ILLUSTRANT DES RELATIONS DE FORMES ET D'INTENSITÉS DIVERSES



Forme et intensité de la relation. Le nuage de points dans un diagramme de dispersion renseigne sur la forme et l'intensité de la relation. Lorsque les points du diagramme sont répartis de façon aléatoire dans l'espace délimité par les deux axes, il n'y a pas de relation entre deux variables. Le nuage de points prend alors une forme arrondie (voir figure 7.12, diagramme *a*). Aussitôt que les points ne semblent pas former un nuage arrondi, c'est qu'il y a une relation. Il y a plusieurs formes de relations et la figure 7.12 en donne quelques exemples. Si les points ont tendance à s'aligner, on dit de la relation qu'elle a une forme **linéaire** (voir diagrammes *b*, *c* et *d*). Si les deux variables varient dans le même sens, c'est-à-dire Y augmente lorsque X augmente, la relation est **positive** (voir diagrammes *b* et *c*). Dans le cas contraire, la relation est **négative** (voir diagramme *d*). Lorsque les points du nuage ne sont pas alignés, la relation n'est pas linéaire et peut alors prendre de multiples formes, par exemple une relation **en courbe** (voir diagrammes *e* et *f*).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

L'intensité d'une relation est reconnaissable dans un diagramme de dispersion à la largeur du nuage de points, ou à sa dispersion de part et d'autre de la ligne imaginaire représentant sa forme. Une relation **forte ou intense** est reconnaissable à un nuage étroit (voir diagrammes *b*, *d* et *e*). Une relation **parfaite** serait constituée d'un ensemble de points formant une ligne et non un nuage. Réciproquement, un nuage dispersé est indicateur d'une relation **peu intense** (voir diagrammes *c* et *f*). Une relation nulle donne un nuage parfaitement arrondi où l'éparpillement des points est à son maximum (voir diagramme *a*).

Revenons maintenant à la figure 7.11 afin de poursuivre l'analyse de la relation entre la moyenne cumulative universitaire et le résultat aux examens de l'Ordre des comptables agréés. On constate qu'il n'y a pas de points dans le coin supérieur gauche ni dans le coin inférieur droit ; le nuage n'a pas une forme arrondie, signe que la dispersion des points n'est pas aléatoire et, conséquemment, qu'il y a une relation entre les deux variables. Voyons de quelle forme est cette relation. Les points forment grossièrement une ellipse allant du coin inférieur gauche au coin supérieur droit et ils ont tendance à se disperser de part et d'autre d'une ligne droite imaginaire. On peut donc dire que la relation est approximativement de forme linéaire. De plus, comme les deux variables varient dans le même sens, la relation est qualifiée de positive. Ainsi, à chaque valeur de la moyenne universitaire correspond un résultat moyen aux examens de l'Ordre, de sorte que plus la moyenne cumulative est élevée, plus le résultat aux examens l'est également.

Une relation autorise la prédiction. Lorsqu'il y a une relation entre deux variables, il est possible de prédire, avec une certaine marge d'erreur, la valeur d'une variable à partir de la valeur de l'autre variable. Par exemple, sachant que la note de passage aux examens de l'Ordre est de 240, combien un diplômé doit-il avoir de moyenne cumulative universitaire pour être certain de réussir les examens de l'Ordre ? Parmi les étudiants de cette promotion de 1980, tous les candidats dont la moyenne universitaire était de 3,4 et plus ont réussi, alors que c'est l'inverse pour ceux qui avaient une moyenne de 2,3 et moins. Quant aux diplômés qui avaient des moyennes entre ces deux valeurs, certains ont réussi et d'autres pas, mais les plus forts au baccalauréat ont eu tendance à obtenir un meilleur taux de succès aux examens de l'Ordre.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Comme la relation n'est pas parfaite, il y a bien des **exceptions**. Par exemple, un diplômé qui avait une moyenne universitaire très faible de 2,4 a tout de même réussi les examens avec 250 comme résultat (voir diplômé n° 12), alors qu'un autre diplômé, avec une moyenne très forte de 3,2, n'a obtenu que 225 aux examens et a échoué (voir diplômé n° 37). Les prédictions sont imprécises et elles doivent être formulées en termes d'intervalle, et non pas en valeurs exactes. Ainsi, un diplômé dont la moyenne universitaire est de 2,5 obtient généralement un résultat aux examens de l'Ordre entre 190 et 265 environ. Naturellement, plus la relation entre les deux variables est intense, plus la prédiction pourra être précise.

Extrapoler les prédictions à d'autres échantillons. À partir de ces données, le directeur du programme de sciences comptables pourra-t-il mieux conseiller ses étudiants désireux de s'inscrire aux examens de l'Ordre ? Autrement dit, est-il permis d'extrapoler ces données aux finissants des années subséquentes qui n'ont pas encore subi les examens de l'Ordre ou qui n'ont même pas encore terminé leur baccalauréat ? Faire des prédictions par rapport à d'autres échantillons que celui de 1980 exige la réunion de certaines conditions. Premièrement, il faut que le présent échantillon soit **représentatif** de l'ensemble des finissants pour lesquels la prédiction sera faite. Deuxièmement, les **conditions de la situation** qui prévalent pour l'échantillon de l'étude, soit la promotion de 1980, doivent être les mêmes que pour la situation de prédiction. Par exemple, si tous les étudiants prennent connaissance de ces données, la situation de prédiction n'est pas la même que celle qui prévalait pour les finissants de l'étude, car ces derniers ne pouvaient pas connaître la relation exacte entre leur moyenne universitaire et leurs chances de succès aux examens de l'Ordre. La connaissance de cette relation pourrait affecter l'attitude des candidats aux examens de multiples façons : nervosité accrue pour les finissants ayant une moyenne universitaire faible, témérité pour ceux dont la moyenne est élevée, etc. Cette question fort importante de la généralisation des résultats est abordée au chapitre 3, à la section « Double validation ».

COEFFICIENTS DE CORRÉLATION (r) ET DE DÉTERMINATION (r^2)

Coefficient de corrélation. Le diagramme de dispersion permet d'observer la forme et l'intensité de la relation entre deux variables. Cependant, lorsqu'il s'agit de quantifier précisément l'intensité de

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

cette relation, il vaut mieux avoir recours à un indice statistique. Le coefficient produit-moment de Pearson, appelé communément coefficient de corrélation ou simplement corrélation, est l'un des indices les plus utilisés, car il présente l'avantage de pouvoir mesurer globalement l'intensité de la relation en un seul chiffre. Toutefois, il comporte un inconvénient de taille : la corrélation produit-moment de Pearson ne peut s'appliquer que si la forme de la relation est linéaire, sinon la corrélation fournira une quantification erronée de la relation, de sorte qu'elle aura systématiquement tendance à sous-estimer la relation réelle[6]. Il faut donc toujours vérifier si la relation entre les deux variables est de forme linéaire avant de calculer la corrélation. Cette vérification peut être faite par l'examen visuel du diagramme de dispersion.

Le coefficient de corrélation, dont le symbole est *r*, provient simplement de la moyenne des produits des deux ensembles de données traduites en notes standard (Dunnette, 1966 ; Schneider et Schmitt, 1986), comme l'exprime la formule suivante[7] :

Coefficient de corrélation linéaire $(r) = \dfrac{\Sigma\, Z_{x_i}\, Z_{y_i}}{N}$

où Z_{x_i} : note standard de chaque observation à la variable X

Z_{y_i} : note standard de chaque observation à la variable Y

Σ : somme de tous les $Z_{x_i}\, Z_{y_i}$

N : nombre d'observations $X_i\, Y_i$

Le calcul de la corrélation pour les 46 finissants entre leur moyenne universitaire et leur résultat aux examens de l'Ordre des comptables agréés est illustré au tableau 7.7 (voir dernière colonne). Le coefficient de corrélation, dont la valeur est de 0,76, provient de

6. Le cas des relations non linéaires est abordé à la fin de la présente section.
7. Cette formule, simple du point de vue théorique, permet de comprendre la notion de corrélation. Signalons qu'il existe plusieurs autres formules, plus pratiques, pour calculer le coefficient.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

la somme des produits des deux variables transformées en notes standard Z (34,96), somme que l'on divise ensuite par le nombre de sujets (46)[8].

Une valeur située entre –1,00 et +1,00. Est-ce que cette corrélation estimée à 0,76 traduit une relation de forte intensité ? En théorie, il y a trois façons de **juger l'ampleur d'une corrélation**. La première façon consiste à apprécier directement la valeur de la corrélation, sachant qu'un coefficient de corrélation est un indice dont la valeur varie entre –1,00 et +1,00 ; le signe indique le sens de la relation alors que le chiffre en traduit l'intensité. Lorsque les notes standard d'une variable correspondent exactement à la grandeur et au signe des notes standard de l'autre variable, la corrélation est de 1,00 : c'est une relation parfaite et positive. Lorsque les notes standard d'une variable correspondent exactement à la grandeur des notes standard de l'autre variable, mais que les signes sont opposés, la corrélation est de –1,00 : c'est une relation inverse et parfaite. Lorsqu'il n'y a aucune liaison entre les notes standard correspondantes, la corrélation vaut 0,00 et indique que la relation est inexistante, comme dans le cas du hasard. Quelques exemples de diagrammes de dispersion et de la valeur approximative des coefficients de corrélation correspondants sont présentés à la figure 7.13.

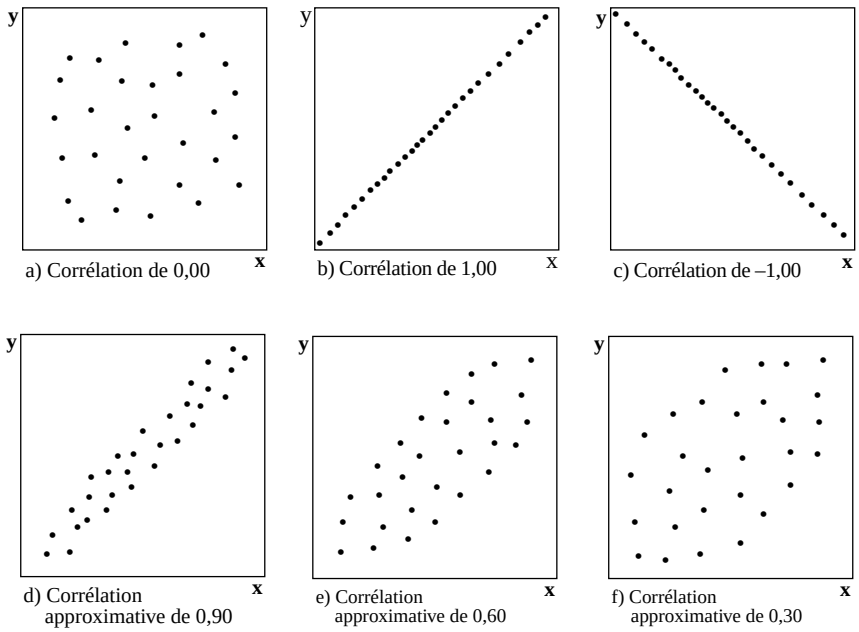
Coefficient de détermination. La valeur des coefficients de corrélation peut être trompeuse pour quiconque est moins familier avec ce genre d'indices. En effet, il faut savoir que les valeurs des coefficients de corrélation sur l'échelle de 0 à 1,00 ne sont pas équidistantes, de sorte que la différence entre deux valeurs s'accroît à mesure que la corrélation augmente et que le point milieu de l'échelle n'est pas 0,50 mais environ 0,71. Par exemple, l'écart de 50 unités entre une corrélation de 0,00 et de 0,50 est sensiblement équivalent à celui de 13 unités entre une corrélation de 0,87 et de 1,00. Heureusement, il est facile de contourner cette difficulté en compilant le coefficient de détermination, qui est simplement l'indice de corrélation élevé au carré. Le coefficient de détermination est normalement exprimé en un pourcentage et la distance entre chaque unité est égale tout au

8. Avec les diverses formules servant à calculer la corrélation directement à partir des notes brutes, il n'est pas nécessaire de convertir les notes brutes en notes standard (Z) : les formules se chargent de cette transformation.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 7.13
DIAGRAMMES DE DISPERSION ILLUSTRANT DIFFÉRENTS COEFFICIENTS DE CORRÉLATION



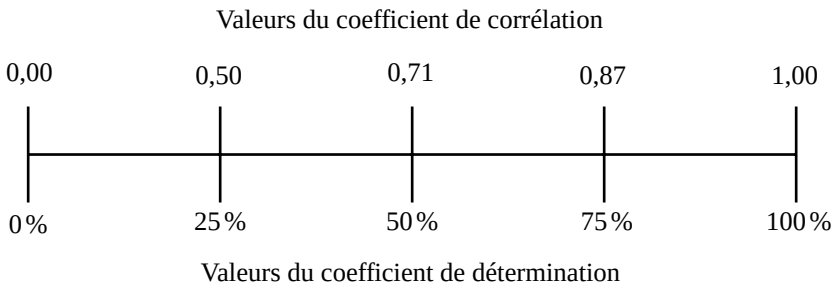
long de cette échelle exprimée en pourcentage. La relation entre les valeurs du coefficient de corrélation et celles du coefficient de détermination est illustrée à la figure 7.14.

Voici la deuxième façon de juger l'ampleur d'une corrélation. Le coefficient de détermination s'interprète comme un pourcentage «d'explication» de la variable dépendante Y par la variable indépendante X, ou réciproquement[9]. Par exemple, la corrélation obtenue avec les 46 candidats aux examens de l'Ordre des comptables agréés est mise au carré pour donner 58 % (soit $0{,}76 \times 0{,}76 = 0{,}58$). Cet indice révèle que la moyenne cumulative universitaire prédit 58 % des résultats aux examens de l'Ordre des comptables agréés.

9. Il serait plus précis de parler en termes de pourcentage de la variation totale d'une variable qui est expliquée par l'autre variable (Baillargeon, 1989).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 7.14
RELATION ENTRE LES VALEURS DU COEFFICIENT DE CORRÉLATION ET CELLES DU COEFFICIENT DE DÉTERMINATION



Test statistique et table des valeurs* r.** La troisième façon de juger si la valeur d'une corrélation est importante réside dans l'utilisation d'un **test statistique**. Un tel test permet de vérifier dans quelle mesure la corrélation observée peut être le fruit du hasard, simplement parce que la corrélation a été obtenue auprès d'un échantillon au lieu de toute la population. Dans l'exemple précédent, la corrélation compilée auprès de l'échantillon de 46 finissants est de 0,76. Voyons la **table des valeurs *r (voir appendice A, tableau A.2). Au niveau du nombre 45 dans la colonne « Taille de l'échantillon (N) », on peut lire les valeurs 0,29 dans la colonne « 0,05 » et 0,38 dans la colonne « 0,01 » ; cela signifie que, dans un échantillon de 45 sujets, il y a 0,05 chance (ou 5 %) d'obtenir par hasard une corrélation de 0,29 et plus, et 0,01 chance (ou 1 %) d'obtenir par hasard une corrélation de 0,38 et plus.

En d'autres mots, si deux variables quelconques sans aucune relation entre elles sont mesurées auprès d'un échantillon de 45 sujets (p. ex., le numéro d'assurance sociale et le poids), il y a tout de même une certaine probabilité d'observer une corrélation tout simplement due au hasard, comme une donne de quatre as en jouant au poker ; grâce à la table des valeurs *r*, il est possible de connaître cette probabilité. Pour un échantillon de 45 observations, la probabilité est de 0,05 pour une corrélation de 0,29 et plus, et de 0,01 pour une corrélation de 0,38 et plus. Dans l'exemple des comptables agréés, la corrélation est de 0,76. Comme cette valeur est nettement au-dessus de 0,38, la probabilité d'avoir obtenu 0,76 est certainement inférieure

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

à 1 %. Par conséquent, il peut être affirmé, avec 1 % de chances de se tromper, que la relation observée entre les deux variables n'est pas due au hasard et qu'elle existe réellement[10].

Insuffisance du cadre statistique. Voilà trois façons, ou plutôt trois perspectives, d'aborder la question de l'ampleur d'une corrélation entre deux variables. Est-ce que la corrélation de 0,76 traduit une relation importante ? Quoique valables en théorie, les trois approches présentées ci-dessus ne permettent pas de répondre à cette question de manière complète. Les statistiques à elles seules ne réussiront jamais à répondre à une telle question ; il faut en plus une connaissance approfondie du contexte pour parvenir à donner un sens aux divers indices statistiques compilés (Tziner, Jeanrie et Cusson, 1993). Par exemple, quel niveau de corrélation a été obtenu dans d'autres situations semblables ? Par rapport aux connaissances et aux théories existantes, quelle doit être l'intensité de la relation entre ces deux variables ? Il faudrait aussi examiner la méthodologie suivie lors de la réalisation de l'étude, l'échantillon, la nature des mesures et la façon avec laquelle elles ont été obtenues, etc. Ces divers éléments contextuels sont pris en considération lorsqu'est abordée la question du

10. Le recours à un test statistique exige la maîtrise des notions d'erreur de type I et de type II, ainsi que leur interrelation avec la taille de l'échantillon et la valeur de la corrélation observée. L'erreur de type I consiste à conclure qu'il y a une corrélation entre les deux variables alors qu'en réalité cette corrélation n'existe pas ; la corrélation observée étant, dans ce cas, le fruit du hasard. L'erreur de type II est l'erreur inverse. Alors que la relation est réelle entre les deux variables, le test nous amène à conclure qu'elle n'existe pas. La puissance d'un test statistique de signification (soit la capacité d'éviter l'erreur de type II) augmente avec la taille de l'échantillon, alors qu'elle décroît à mesure que la corrélation observée se rapproche de 0,00 et que la probabilité de l'erreur de type I diminue. L'explication est simple : plus la taille de l'échantillon diminue, plus les corrélations peuvent être affectées par le hasard. Conséquemment, pour maintenir constante la probabilité α de commettre l'erreur de type I, il faut augmenter la valeur critique de *r* au test statistique. Cependant, cette compensation par la majoration de la valeur critique de *r* a pour effet d'entraîner une baisse de puissance du test statistique, de sorte qu'une relation modérée, mais réelle, risque de passer pour le fruit du hasard et de se voir ainsi rejetée. Ce domaine de la statistique est trop vaste pour être expliqué de manière satisfaisante dans cet ouvrage. Pour appliquer et interpréter de manière rigoureuse un test statistique, il faut absolument s'en remettre à un ouvrage spécialisé.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

niveau de validité critériée souhaitable pour un instrument de sélection du personnel (voir chapitre 3, la section « Appréciation de la grandeur du coefficient de validité »).

Limites du coefficient de corrélation. Le coefficient de corrélation, même s'il est un des indices les plus utilisés, n'offre pas que des avantages. Dunnette (1966) rappelle avec insistance deux limites particulièrement importantes sur le plan pratique. La première limite du coefficient de corrélation est qu'il ne peut pas fournir une estimation valable de la relation entre deux variables si la forme de la relation n'est pas linéaire. Le coefficient de corrélation est un indice basé sur le fait que les données sont en relation linéaire. Si la forme de la relation n'est pas linéaire, le calcul de la corrélation ne permettra pas de saisir la nature de la relation ni d'en apprécier l'intensité ; la corrélation obtenue indiquera une relation plus faible que celle qui existe en réalité.

En guise d'exemple, la figure 7.15 montre la relation, chez 45 femmes travaillant en usine, entre un test d'aptitude verbale passé lors de la sélection et l'évaluation de leur rendement deux ans plus tard. En apparence, il ne semble pas y avoir de relation entre le test d'aptitude verbale et le rendement chez ces employées (voir diagramme de gauche). D'ailleurs, le calcul de la corrélation donne un indice avoisinant 0,00. Cependant, en y regardant de plus près, on peut déceler une légère relation non linéaire en forme de « U » inversé. En effet, en divisant le diagramme en tranches verticales, en calculant la moyenne du rendement pour chacune de ces tranches, en indiquant cette moyenne par un *x* et en joignant par une ligne ces *x*, on obtient une ligne courbe (voir diagramme de droite). Cette relation pourrait traduire une situation où des employées surqualifiées connaissent une diminution de leur motivation parce qu'elles ne sentent pas que leur potentiel est utilisé pleinement. Comme on pouvait s'y attendre, l'intensité de cette relation est très faible[11].

11. Dans le cas d'une relation non linéaire, le coefficient de corrélation « êta » est plus approprié (Schneider et Schmitt, 1986).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 7.15
DIAGRAMMES DE DISPERSION POUR 45 EMPLOYÉES EN USINE ENTRE LEUR SCORE À UN TEST D'APTITUDE VERBALE ET LEUR ÉVALUATION DU RENDEMENT

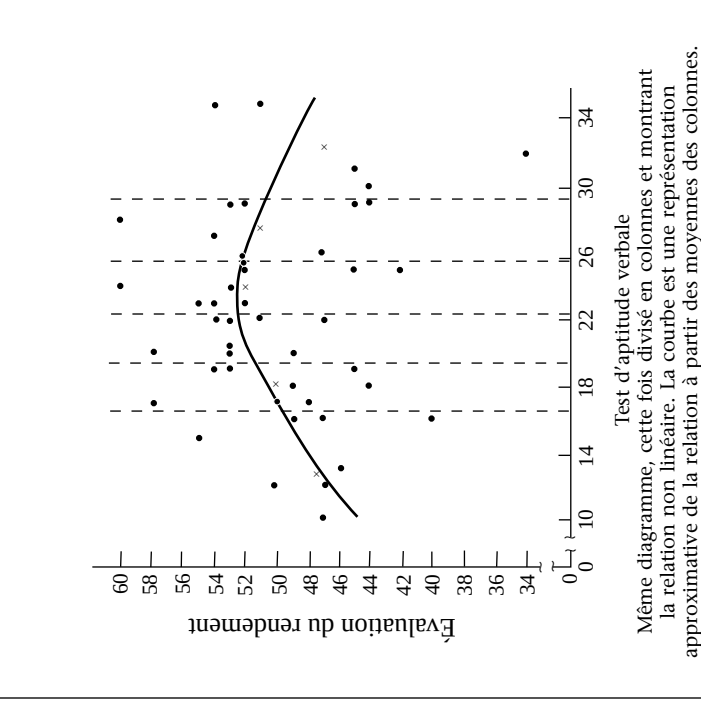


Diagramme où la corrélation linéaire est zéro.

Même diagramme, cette fois divisé en colonnes et montrant la relation non linéaire. La courbe est une représentation approximative de la relation à partir des moyennes des colonnes.

Source : Guion (1965, p. 133-134).

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

La deuxième limite du coefficient de corrélation est qu'il n'estime pas d'emblée la relation entre des **sous-ensembles** de données. Comme une moyenne, il ne peut révéler les particularités. Ainsi, s'il se trouve deux sous-ensembles d'observations pour lesquels les relations sont différentes, le calcul de la corrélation ne pourra pas révéler la nature ni l'intensité de ces relations.

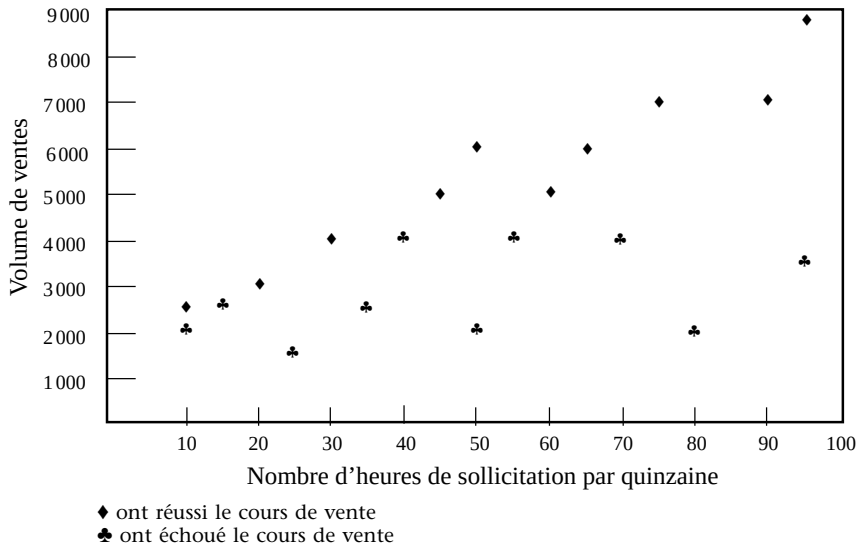
Prenons l'exemple fictif d'un échantillon de représentants commerciaux dont le volume de ventes est mis en relation avec leur nombre d'heures passées à solliciter les clients (voir figure 7.16). L'examen du diagramme de dispersion révèle que la relation n'est pas la même selon que les représentants ont réussi ou échoué un cours spécialisé dans la vente. Le calcul de la corrélation pour l'échantillon total donne un coefficient de 0,61, laissant supposer une relation d'intensité modérée pour l'ensemble des représentants. Pourtant, la réalité est fort différente : il y a une relation presque parfaite entre le volume des ventes et le nombre d'heures de sollicitation pour les représentants ayant réussi leur cours de vente ($r = 0{,}96$), alors que la relation est beaucoup plus faible pour ceux qui ont échoué ($r = 0{,}43$). Bref, le coefficient de corrélation, calculé sans avoir au préalable examiné le diagramme de dispersion, peut parfois dissimuler beaucoup plus qu'il ne révèle (Dunnette, 1966). En statistique, la prudence est toujours de rigueur.

DROITE DE RÉGRESSION ET ERREUR TYPE DE L'ESTIMATION (S_{ee})

À l'aide du diagramme de dispersion, nous avons vu qu'il est possible de prédire, avec une certaine marge d'erreur, la valeur d'une variable à partir de la valeur de l'autre variable. De plus, dans le cas où cette relation est linéaire, nous savons que l'utilisation du coefficient de corrélation permet d'en évaluer l'intensité. Il est possible de poursuivre l'analyse de la relation entre deux variables en ayant recours à d'autres outils statistiques fort intéressants, comme la régression. Cette technique permet, si une relation est effectivement observée entre deux variables, de prédire la valeur d'une variable à partir de l'autre, puis d'évaluer précisément la marge d'erreur associée à cette prédiction. La régression est un domaine qui comporte plusieurs concepts statistiques. Pour notre présente démarche, qui consiste à introduire quelques éléments statistiques essentiels à la gestion des ressources humaines, deux concepts seulement retiendront l'attention :

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 7.16
DIAGRAMME DE DISPERSION DU VOLUME DE VENTES DE 20 REPRÉSENTANTS COMMERCIAUX EN FONCTION DE LEUR NOMBRE D'HEURES DE SOLLICITATION PAR QUINZAINE, COMPTE TENU DE LEUR NIVEAU D'APTITUDES VERBALES

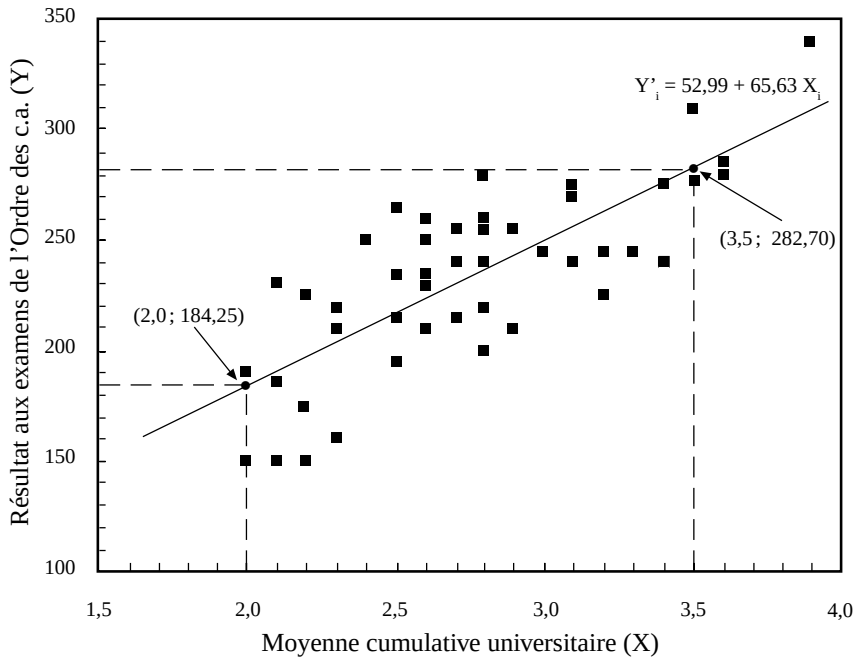


la droite de régression et l'erreur type de l'estimation ($S_{ée}$). De plus, l'exposé se limitera au modèle linéaire simple, c'est-à-dire lorsque la relation est de forme linéaire et implique une seule variable indépendante.

Droite de régression. Les données provenant de la figure 7.11, relatives aux diplômés en comptabilité, ont été reprises à la figure 7.17. Dans cette nouvelle figure, on remarque qu'une droite oblique a été tracée en travers du nuage de points. Cette ligne, appelée droite de régression, a été calculée de manière à constituer la meilleure représentation linéaire possible de la relation entre la moyenne cumulative universitaire et les résultats à l'examen de l'Ordre des comptables agréés. Son emplacement a été déterminé pour réduire au minimum la distance entre cette droite et les divers points du diagramme.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 7.17
DROITE DE RÉGRESSION ENTRE LA MOYENNE CUMULATIVE UNIVERSITAIRE ET LE RÉSULTAT AUX EXAMENS DE L'ORDRE DES COMPTABLES AGRÉÉS POUR 46 CANDIDATS



Mathématiquement, une ligne peut être représentée par la formule suivante :

Droite de régression $(Y'_i) = a + b\ X_i$
où Y'_i : valeur estimée de la variable Y pour X_i
a : ordonnée à l'origine de la droite de régression
b : pente de la droite de régression
X_i : chaque observation à la variable X

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

Dans l'exemple précédent (figure 7.17), la droite de régression est définie par l'équation $Y'_i = 52{,}99 + 65{,}63\ X_i$, où $a = 52{,}99$ et $b = 65{,}63$. Les formules servant à calculer les valeurs de a et de b sont les suivantes (Baillargeon, 1989, p. 442)[12] :

Pente $(b) = \dfrac{\Sigma X_i\, Y_i - \dfrac{(\Sigma X_i)(\Sigma Y_i)}{N}}{\Sigma X_i^2 - \dfrac{(\Sigma X_i)^2}{N}}$

où X_i : chaque observation de la variable X

Y_i : chaque observation de la variable Y

N : nombre d'observations $X_i\ Y_i$

Ordonnée à l'origine $(a) = M_y - b\ M_x$

où M_y : moyenne des observations de la variable Y

b : pente de la droite de régression

M_x : moyenne des observations de la variable X

Pour tracer la droite de régression, on calcule, à l'aide de l'équation ci-dessus, deux valeurs de Y'_i correspondant à deux valeurs de X_i, ce qui permet de placer ces deux points (X_i et Y'_i) dans le diagramme, puis on les relie par une ligne droite. Les valeurs de Y'_i pour $X_i = 2{,}0$ et pour $X_i = 3{,}5$ ont été calculées :

$Y'_i = a + b\ X_i$ où $a = 52{,}99$ et $b = 65{,}63$

$Y'_{2,0} = 52{,}99 + (65{,}63)\ (2{,}0) = 184{,}25$

$Y'_{3,5} = 52{,}99 + (65{,}63)\ (3{,}5) = 282{,}70$

Ces deux points (2,0 ; 184,25) et (3,5 ; 282,70) ont ensuite été reportés sur le diagramme de dispersion, pour donner la droite de régression de la figure 7.17.

12. Les données de l'exemple proviennent du tableau 7.7, pour lesquelles les calculs préliminaires suivants ont été effectués : $\Sigma X_i = 126{,}8$, $\Sigma X_i^2 = 359{,}9$, $\Sigma Y_i = 10\ 770$, $\Sigma Y_i^2 = 2\ 596\ 600$ et $\Sigma X_i\ Y_i = 30\ 368{,}5$. Rappelons que $N = 46$.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca

Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

L'équation ou la droite de régression sert à faire des prédictions. Supposons un diplômé qui a obtenu 2,9 comme moyenne cumulative universitaire pour son baccalauréat en comptabilité. Supposons également que les conditions de la situation (programme d'études, force des étudiants, examens de l'Ordre des comptables agréés, etc.) sont équivalentes à celles qui prévalaient pour les 46 finissants de 1980. Alors, l'équation de régression permet d'estimer le résultat attendu aux examens de l'Ordre des comptables agréés pour ce candidat. Voyons les calculs.

Si $X_i = 2,9$ alors $Y'_{2,9} = 52,99 + (65,63)\ (2,9) = 243,32$.

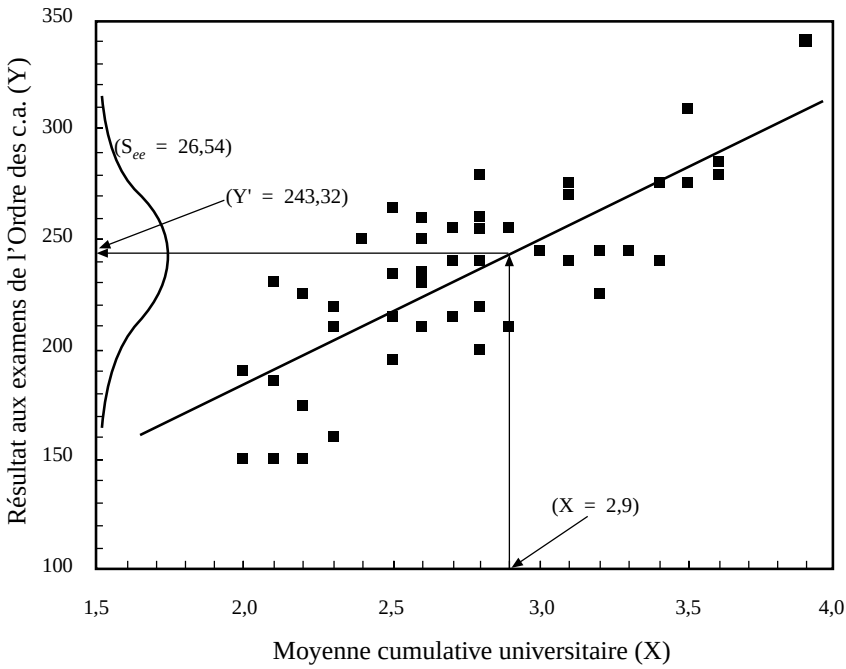
Une autre méthode pour estimer le résultat attendu consiste simplement à recourir directement à la droite de régression qui a été tracée sur le diagramme de dispersion. Reprenons l'exemple du candidat ayant obtenu 2,9 de moyenne cumulative (voir figure 7.18). On n'a qu'à localiser cette valeur de X = 2,9 sur l'axe horizontal, à tracer une ligne perpendiculaire jusqu'à la droite de régression, puis, à partir de ce point d'intersection, à tracer une ligne parallèle à l'axe horizontal jusqu'à l'axe vertical. Ce dernier point d'arrivée correspond à la valeur attendue aux examens de l'Ordre des comptables agréés, soit environ 243.

Ainsi, en calculant l'équation de régression ou en utilisant directement la droite de régression, on peut prédire que ce candidat obtiendra un résultat de 243 aux examens de l'Ordre des comptables agréés. Il est évident qu'une telle prédiction ne peut être parfaite, étant donné que l'intensité de la relation entre les deux variables (soit la moyenne cumulative et les résultats aux examens de l'Ordre des comptables agréés) ne l'est pas. La dispersion du nuage de points autour de la droite de régression et le coefficient de corrélation de 0,76 calculé précédemment traduisent cette imperfection. D'ailleurs, deux candidats de l'échantillon avaient effectivement obtenu une moyenne cumulative universitaire de 2,9 et leur résultat aux examens de l'Ordre des comptables agréés n'a pas été 243, mais bien 210 et 255. Donc, les prévisions faites à l'aide de la régression ne sont pas parfaites et ne peuvent être meilleures que l'intensité de la relation observée entre les deux variables en cause.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Figure 7.18
ESTIMATION DU RÉSULTAT AUX EXAMENS DE L'ORDRE DES COMPTABLES AGRÉÉS À PARTIR DE LA MOYENNE CUMULATIVE UNIVERSITAIRE POUR 46 DIPLÔMÉS



Erreur type de l'estimation (S_{ee}). Alors, si la prédiction n'est pas parfaite, il serait important d'en connaître la marge d'erreur. Par exemple, est-ce possible que le candidat qui a une moyenne cumulative universitaire de 2,9 obtienne un résultat de 300 et plus aux examens de l'Ordre des comptables agréés ? Quelle est la probabilité qu'une telle situation se produise ? L'examen du diagramme de dispersion fournit une certaine indication de la marge d'erreur associée à cette prédiction. Cependant, ces deux seuls diplômés constituent un échantillon bien petit et ne permettent pas d'estimer précisément la dispersion de tous les résultats attendus possibles de Y (examens de l'Ordre des comptables agréés) pour une valeur obtenue de X

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

(moyenne cumulative universitaire). Pour connaître la marge d'erreur, il vaut mieux s'en remettre à l'erreur type de l'estimation (S_{ee}), dont la formule est la suivante (Guion, 1965 ; Arvey et Faley, 1988) :

Erreur type de l'estimation $(S_{ee}) = S_y \sqrt{1 - r_{xy}}$
où S_y : écart type des observations à la variable Y
r_{xy} : corrélation observée entre la variable X et la variable Y

Il est intéressant de noter que la corrélation fait partie de la formule, de sorte que l'erreur type de l'estimation variera inversement à l'intensité de la relation entre les deux variables X et Y. Plus la corrélation est élevée, plus l'erreur type de l'estimation sera petite. Si la corrélation est parfaite ($r = 1{,}00$), alors l'erreur type de l'estimation sera nulle ($S_{ee} = 0$) ; si la corrélation est nulle ($r = 0{,}00$), alors l'erreur type de l'estimation sera à son maximum ($S_{ee} = S_y$).

L'erreur type de l'estimation est considérée comme l'écart type de la distribution des erreurs de prédiction qui existe autour de Y'_i (estimation de la variable Y effectuée à partir de la variable X). Comme ces erreurs de prédiction sont présumées se **distribuer normalement** et être constantes pour toutes valeurs de X, on peut s'en remettre aux propriétés de la courbe normale pour en interpréter la signification (voir figure 7.6). Pour toutes prévisions de la variable Y, il y a 68,26 % de chances que l'erreur ne dépasse pas une erreur type de l'estimation. La probabilité est de 68,26 % que le score qui sera réellement obtenu (Y) se retrouve autour de la valeur estimée par régression (Y'), quelque part entre une S_{ee} en dessous de cette prédiction et une S_{ee} au-dessus. Réciproquement, la probabilité d'une erreur supérieure à une S_{ee} en plus ou en moins de la valeur attendue par régression est de 31,74 % (soit 100 % – 68,26 %).

Revenons à l'exemple de la figure 7.18 dont les données proviennent du tableau 7.7. L'erreur type de l'estimation de Y pour ces 46 candidats est de 26,54. Se référant à la répartition des observations dans une distribution normale (figure 7.6), on peut affirmer qu'un candidat ayant une moyenne cumulative de 2,9 obtiendra un résultat

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N
Tous droits de reproduction, de traduction ou d'adaptation réservés

aux examens d'environ 243, plus ou moins une marge d'erreur qui se distribue normalement avec un écart type, appelé erreur type de l'estimation, de 26,54 points. On peut voir cette distribution le long de l'axe vertical et centrée autour de la valeur 243,32 qui avait été estimée[13].

L'application de la régression et l'analyse de ses résultats exigent la maîtrise de plusieurs concepts statistiques, dont la présentation, nous le rappelons, dépasse le cadre de ce livre. Le lecteur désireux d'appliquer de manière pleinement satisfaisante cette technique devra consulter des ouvrages statistiques.

Des outils au service de l'intelligence. Aussi puissants que soient les outils d'analyses statistiques, ils n'en demeurent pas moins des outils ; rien ne peut se substituer à la perspicacité et au jugement du professionnel. La clé d'une décision tout à fait satisfaisante demeure la connaissance approfondie des phénomènes, des particularités de la situation et de l'importance qu'on leur accorde.

13. Pour les personnes déjà familières avec l'analyse de régression, il convient de souligner que la notion d'erreur type de l'estimation est semblable à l'écart type de la distribution des résidus. Par conséquent, cet indice ne tient pas compte de l'erreur d'échantillonnage. Pour une estimation de la marge d'erreur qui s'applique réellement à l'ensemble de la population, il faudrait utiliser l'écart type de l'erreur de prévision (Baillargeon, 1989, p. 475). Ainsi, pour l'exemple des candidats en comptabilité, la marge d'erreur pour la prédiction aux examens de l'Ordre des comptables agréés applicable à toute la population des diplômés ne serait pas 26,54 (soit S_{ee} actuelle), mais 27,31. Une différence aussi minime n'aurait pas de conséquences en gestion des ressources humaines. Notons en terminant que cette dernière marge d'erreur s'applique uniquement lorsque la moyenne cumulative universitaire observée est de 2,9. L'écart type de l'erreur de prévision n'est pas constante pour toutes les valeurs de X, de sorte qu'elle augmente légèrement à mesure que la valeur de X s'éloigne de la moyenne des observations pour cette variable.

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

APPENDICE A
TABLES STATISTIQUES

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau A.1
TABLE DE LA LOI NORMALE CENTRÉE RÉDUITE

Z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0000	0040	0080	0120	0159	0199	0239	0279	0319	0359
0,1	0398	0438	0478	0517	0557	0596	0636	0675	0714	0753
0,2	0793	0832	0871	0910	0948	0987	1026	1064	1103	1141
0,3	1179	1217	1255	1293	1331	1368	1406	1443	1480	1517
0,4	1554	1591	1628	1664	1700	1736	1772	1808	1844	1879
0,5	1915	1950	1985	2019	2054	2088	2123	2157	2190	2224
0,6	2257	2291	2324	2357	2389	2422	2454	2486	2518	2549
0,7	2580	2612	2642	2673	2704	2734	2764	2794	2823	2852
0,8	2881	2910	2939	2967	2995	3023	3051	3078	3106	3133
0,9	3159	3186	3212	3238	3264	3289	3315	3340	3365	3389
1,0	3413	3438	3461	3485	3508	3531	3554	3577	3599	3621
1,1	3643	3665	3686	3718	3729	3749	3770	3790	3810	3830
1,2	3849	3869	3888	3907	3925	3944	3962	3980	3997	4015
1,3	4032	4049	4066	4083	4099	4115	4131	4147	4162	4177
1,4	4192	4207	4222	4236	4251	4265	4279	4292	4306	4319
1,5	4332	4345	4357	4370	4382	4394	4406	4418	4430	4441
1,6	4452	4463	4474	4485	4495	4505	4515	4525	4535	4545
1,7	4554	4564	4573	4582	4591	4599	4608	4616	4625	4633
1,8	4641	4649	4656	4664	4671	4678	4686	4693	4699	4706
1,9	4713	4719	4726	4732	4738	4744	4750	4758	4762	4767
2,0	4773	4778	4783	4788	4793	4789	4803	4808	4812	4817
2,1	4821	4826	4830	4834	4838	4842	4846	4850	4854	4857
2,2	4861	4865	4868	4871	4875	4878	4881	4884	4887	4890
2,3	4893	4896	4898	4901	4904	4906	4090	4911	4913	4916
2,4	4918	4920	4922	4925	4927	4929	4931	4932	4934	4936
2,5	4938	4940	4941	4943	4945	4946	4948	4949	4951	4952
2,6	4953	4955	4956	4957	4959	4960	4961	4962	4963	4964
2,7	4965	4966	4967	4968	4969	4970	4971	4972	4973	4974
2,8	4974	4975	4976	4977	4977	4978	4979	4980	4980	4981
2,9	4981	4982	4983	4984	4984	4984	4985	4985	4986	4986
3,0	4986,5	4987	4987	4988	4988	4988	4989	4989	4989	4990
3,1	4990,0	4991	4991	4991	4992	4992	4992	4992	4993	4993
3,2	4993,129									
3,3	4995,166									
3,4	4996,631									
3,5	4997,674									
3,6	4998,409									
3,7	4998,922									
3,8	4999,277									
3,9	4999,519									
4,0	4999,683									
4,5	4999,966									
5,0	4999,997133									

La loi normale centrée réduite, dont la moyenne (μ) égale 0 et l'écart type (σ) vaut 1, donne la proportion d'observation comprise entre la moyenne et toute valeur de *z*. Les valeurs *z* apparaissent dans la colonne pour les unités et les dixièmes, alors que les centièmes sont dans la ligne supérieure. Les valeurs qui apparaissent dans le corps du tableau sont les proportions d'observations recherchées. Par exemple, la proportion d'observations comprise entre la moyenne 0 et la valeur *z* 1,15, est de 0,3749 ou 37,49 % des cas.
Source : H.O. Rugg (1917), *Statistical Methods Applied to Education*. Boston, Houghton, Mifflin.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau A.2
TABLE DES VALEURS *R* POUR LA CORRÉLATION LINÉAIRE ENTRE DEUX VARIABLES

Taille de l'échantillon (N)	Seuil de probabilité (α)	
	0,05	**0,01**
3	0,98*	1,00*
4	0,95	0,99
5	0,88	0,96
6	0,81	0,92
7	0,75	0,87
8	0,71	0,83
9	0,66	0,80
10	0,63	0,76
11	0,60	0,73
12	0,57	0,71
13	0,55	0,68
14	0,53	0,66
15	0,51	0,64
20	0,44	0,56
25	0,40	0,50
30	0,36	0,46
35	0,33	0,43
40	0,31	0,40
45	0,29	0,38
50	0,27	0,36
70	0,23	0,30
100	0,19	0,25
200	0,15	0,18

***** Valeur de *r* qui est statistiquement significative pour le seuil de probabilité (a) et la taille de l'échantillon (N) correspondants.

Source : R.D. Arvey et R.H. Faley (1988), *Fairness in Selecting Employees* (2e éd. rev.), Reading, MA, Addison-Wesley ; R.W. Beatty et C.E. Schneier (1977), *Personnel Administration : An Experiential Skill-Building Approach,* Reading, MA, Addison-Wesley.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Tableau A.3
TABLE DE LA DISTRIBUTION DE STUDENT

dl	Niveau de signification pour un test unilatéral					
	,10	,05	,025	,01	,005	,0005
	Niveau de signification pour un test bilatéral					
	,20	,10	,05	,02	,01	,001
1	3,078	6,314	12,706	31,821	63,657	636,619
2	1,886	2,920	4,303	6,965	9,925	31,598
3	1,638	2,353	3,182	4,541	5,841	12,941
4	1,533	2,132	2,776	3,747	4,604	8,610
5	1,476	2,015	2,571	3,365	4,032	6,859
6	1,440	1,943	2,447	3,143	3,707	5,959
7	1,415	1,895	2,365	2,998	3,499	5,405
8	1,397	1,860	2,306	2,896	3,355	5,041
9	1,383	1,833	2,262	2,821	3,250	4,781
10	1,372	1,812	2,228	2,764	3,169	4,587
11	1,363	1,796	2,201	2,718	3,106	4,437
12	1,356	1,782	2,179	2,681	3,055	4,318
13	1,350	1,771	2,160	2,650	3,012	4,221
14	1,345	1,761	2,145	2,624	2,977	4,140
15	1,341	1,753	2,131	2,602	2,947	4,073
16	1,337	1,746	2,120	2,583	2,921	4,015
17	1,333	1,740	2,110	2,567	2,898	3,965
18	1,330	1,734	2,101	2,552	2,878	3,922
19	1,328	1,729	2,093	2,539	2,861	3,883
20	1,325	1,725	2,086	2,528	2,845	3,850
21	1,323	1,721	2,080	2,518	2,831	3,819
22	1,321	1,717	2,074	2,508	2,819	3,792
23	1,319	1,714	2,069	2,500	2,807	3,767
24	1,318	1,711	2,064	2,492	2,797	3,745
25	1,316	1,708	2,060	2,485	2,787	3,725
26	1,315	1,706	2,056	2,479	2,779	3,707
27	1,314	1,703	2,052	2,473	2,771	3,690
28	1,313	1,701	2,084	2,467	2,763	3,674
29	1,311	1,699	2,045	2,462	2,756	3,659
30	1,310	1,697	2,042	2,457	2,750	3,646
40	1,303	1,684	2,021	2,423	2,704	3,551
60	1,296	1,671	2,000	2,390	2,660	3,460
120	1,289	1,658	1,980	2,358	2,617	3,373
∞	1,282	1,645	1,960	2,326	2,576	3,291

Les valeurs contenues dans la table sont des valeurs *t* associées à divers seuils α (ou seuils de signification) et degrés de liberté (dl). Par exemple, la valeur *t* pour un seuil α de 0,01 et un degré de liberté de 15 est 2,947 pour un test bilatéral (2,602 pour un test unilatéral). La probabilité d'obtenir par hasard une valeur *t* égale ou supérieure à cette valeur 2,947 de la table est égale au seuil α correspondant, soit 0,01.

Source : R.A. Fisher et F. Yate (1948), *Statistical Tables for Biological, Agricultural and Medical Research*, Edimbourg et Londres, Oliver & Boyd.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

APPENDICE B
LIGNES DIRECTRICES EN MESURE ET ÉVALUATION

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington DC : American Psychological Association.

C'est l'ouvrage de référence le plus complet et probablement le plus utilisé en matière de lignes directrices concernant les instruments de mesure et leurs usages. Rédigés conjointement par l'American Educational Research Association, l'American Psychological Association et le National Council on Measurement in Education, les *Standards* fournissent un cadre de référence pour évaluer les instruments de mesure, leur application et les effets de leur application. De plus, les *Standards* viennent porter assistance lorsque intervient le jugement professionnel basé sur les principes généralement reconnus (*accepted corpus of knowledge*, p. 2). Les membres de l'Ordre professionnel des psychologues du Québec sont assujettis à ces *Standards* par l'article 72 de leur code de déontologie.

Société canadienne de psychologie (1987). *Lignes directrices pour les tests psychologiques et pédagogiques* (en anglais, *Guidelines for educational and psychological testing*).

Il s'agit de l'adaptation canadienne du document *Standards for educational and psychological testing* (version 1985), réalisée sous l'égide de la Société canadienne de psychologie.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

Society for Industrial and Organizational Psychology (1987). *Principles for validation and use of personnel selection procedures* (3e éd., rev.). College Park, MD.

Cet ouvrage contient des lignes directrices adaptées aux pratiques de sélection et de promotion du personnel. Il constitue la position officielle de la Society for Industrial and Organizational Psychology, une division de l'American Psychological Association.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, & Department of Justice (1978). *Adoption by four agencies of uniform guidelines on employee selection procedures.* Federal Register, 43(166), 38290-38315.

Ce document a été élaboré par le gouvernement américain et contient les lignes à suivre en matière d'évaluation et de sélection afin d'éviter les pratiques discriminatoires.

Centre de psychologie du personnel (1990). *L'évaluation des compétences #14 – Les tests dans l'administration publique fédérale*. Ottawa : Commission de la fonction publique du Canada.

Les lignes directrices émises par la Société canadienne de psychologie ont servi de fondement à l'élaboration des présentes lignes directrices, mais dans le contexte plus particulier de l'administration publique fédérale.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

RÉFÉRENCES

AIKEN, L.R. (1988). *Psychological testing* (2e éd. rev.). Boston : Allyn et Bacon.

AMERICAN EDUCATIONAL RESEARCH ASSOCIATION, AMERICAN PSYCHOLOGICAL ASSOCIATION, ET NATIONAL COUNCIL ON MEASUREMENT IN EDUCATION (1985, 1999). *Standards for educational and psychological testing*. Washington, D.C. : American Psychological Association.

ARTHUR Jr, W., DOVERSPIKE, D. et BARRETT, G.V. (1996). Development of a job analysis-based procedure for weighting and combining content-related tests into a single test battery score. *Personnel Psychology, 49*, 971-985.

ARVEY, R.D. et FALEY, R.H. (1988). *Fairness in selecting employees* (2e éd. rev.). Reading, MA : Addison-Wesley.

AUSTIN, J.T. et VILLANOVA, P. (1992). The criterion problem : 1917–1922. *Journal of Applied Psychology, 77*, 936-874.

BAEHR, M.E. (1993). *Predicting success in higher-level positions : A guide to the system for testing and evaluating potential*. Westport, CT : Quorum Books.

BAILLARGEON, G. (1989). Probabilités, statistique et techniques de régression. Trois-Rivières, QC : Les Éditions SMG.

BAILLARGEON, G. et MARTIN, L. (1994). *Méthodes quantitatives et analyse de données. Tome 1 : analyse descriptive*. Trois-Rivières, QC, Canada : Les Éditions SMG.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca

Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

BAKER, B.O., HARDICK, C.D. et PETRINOVICH, L.F. (1966). Weak measurement vs strong statistics : An empirical critique of S. Stevens' proscriptions on statistics. *Educational and Psychological Measurement, 26*, 291-309.

BARRET, G.V., PHILLIPS, J.S. et ALEXANDER, R.A. (1981). Concurrent and predictive validity designs : A critical reanalysis. *Journal of Applied Psychology, 66*, 1-6.

BARRETTE, J. et DURIVAGE, A. (1993). Sélection du personnel : les connaissances tacites permettent-elles de prédire le succès au travail ? Document de travail 93-56. Ottawa : Université d'Ottawa.

BARRICK, M.R. et MOUNT, M.K. (1991). The big five personality dimensions and job performance : A meta-analysis. *Personnel Psychology, 44*, 1-26.

BEATTY, R.W. et SCHNEIDER, C.E. (1977). *Personnel administration : An experiential skill-building approach*. Reading, MA : Addison-Wesley.

BINNING, J.F. et BARRETT, G.V. (1989). Validity of personnel decisions : A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*, 478-494.

BOBKO, P., ROTH, P.L. et POTOSKY, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology, 52*, 561-589.

BORMAN, W.C. (1991). Job behavior, performance, and effectiveness. *In* M.D. Dunnette et L.M. Hough (éd.). *Handbook of industrial and organizational psychology* (2e éd. rev.), *vol. 2* (p. 271-326). Palo Alto, CA : Consulting Psychologists Press.

BOUDREAU, J.W. (1991). Utility analysis for decisions in human resource management. *In* M.D. Dunnette et L.M. Hough (éd.). *Handbook of industrial and organizational psychology* (2e éd. rev.), *vol. 2* (p. 621-745). Palo Alto, CA : Consulting Psychologists Press.

BOMMER, W.H., JOHNSON, J.L., RICH, G.A., PODSAKOFF, P.M. et MACKENZIE, S.B. (1995). On the interchangeability of objective and subjective measures of employee performance : A meta-analysis. *Personnel Psychology, 48*, 587-605.

BOYATZIS, R.E. (1982). *The competent manager*. New York : Wiley.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

BRADBURN, N.M. et SUDMAN, S. (1982). *Asking questions*. San Francisco : Jossey-Bass.

BROGDEN, H.E. (1949). When testing pays off. *Personnel Psychology, 2*, 171-183.

CAMPBELL, J.P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. *In* M.D. Dunnette et L.M. Hough (éd.). *Handbook of industrial and organizational psychology* (2e éd. rev.), *vol. 1* (p. 687-732). Palo Alto, CA : Consulting Psychologists Press.

CAMPBELL, J.P., MCHENRY, J.J. et WISE, L.L. (1990). Modeling job performance in a population of jobs. *Personnel Psychology, 43*, 313-333.

CAMPBELL, J.P., MCCLOY, R.A., OPPLER, S.H. et SAGER, C.E. (1993). A theory of performance. *In* N. Schmitt et W.C. Borman (éd.), *Personnel selection* (p. 35-70). San Francisco : Jossey-Bass.

CARMINES, E.G. et ZELLER, R.A. (1979). *Reliability and validity assessment*. Newbury Park, CA : Sage.

CASCIO, W.F. (1982). *Costing human resources : The financial impact of behavior in organizations*. Boston, MA : Kent.

CASCIO, W.F. (1987). *Applied psychology in personnel management* (3e éd. rev.). Reston, VA : Reston.

CASCIO, W.F. (1993). Assessing the utility of selection decisions : Theoretical and practical considerations. *In* N. Schmitt, W.C. Borman et collab. (éd.). *Personnel selection in organizations* (p. 310-340). San Francisco : Jossey-Bass.

CASCIO, W.F. et RAMOS, R.A. (1986). Development and application of a new method for assessing job performance in behavior/economic terms. *Journal of Applied Psychology, 71*, 20-28.

CASCIO, W.F. et SWEET, D.H. (1989). *Human resource planning, employment, and placement*. Washington, D.C. : The Bureau of National Affairs.

CASCIO, W.F., ALEXANDER, R.A. et BARRETT, G.V. (1988). Setting cutoff scores : legal, psychometric, and professional issues and guidelines. *Personnel Psychology, 41* (1), 1-24.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

CATANO, V.M., CRONSHAW, S.F., WIESNER, W.H., HACKETT, R.D. et MÉTHOT, L.L. (1997). *Recruitment and selection in Canada*. Scarborough, Ont. : ITP Nelson.

CATTIN, P. (1980). Estimating of the predictor power of a regression model. *Journal of Applied Psychology, 65*, 407-414.

CENTRE DE PSYCHOLOGIE DU PERSONNEL (1984). *L'élaboration des tests de connaissances*. Ottawa : Commission de la fonction publique du Canada, Ministère des Approvisionnements et Services Canada.

CENTRE DE PSYCHOLOGIE DU PERSONNEL (1987). *In-Basket exercise 810 (IBE 810) – Technical manual*. Ottawa : Commission de la fonction publique du Canada.

CHAGNON, Y. (1991). Conceptualisation de la culture organisationnelle et élaboration d'un instrument de mesure. Thèse de doctorat inédite, Université de Montréal.

CONWAY, J.M., JAKO, R.A. et GOODMAN, D.F. (1995). A meta-analysis of interrater and internal consistency reliability of selection interviews. *Journal of Applied Psychology, 80*, 565-579.

COOK, M. (1988). *Personnel selection and productivity*. Chichester, Angleterre : John Wiley.

COOPER, D.R. et SCHINDLER, P.S. (1998). *Business research methods* (6e éd. rev.). New York : Irwin/McGraw-Hill.

COWARD, W.M. et SACKETT, P.R. (1990). Linearity of ability-performance relationship : A reconfirmation. *Journal of Applied Psychology, 75*, (3) 297-300.

CRONBACH, L.J. (1970). *Essentials of psychological testing* (3e éd. rev.). New York : Harper et Row.

CRONBACH, L.J. (1971). Test validation. *In* R.L. Thorndike (éd.), Educational measurement (2e éd. rev.). Washington, D.C. : American Council on Education.

CRONBACH, L.J. (1990). *Essentials of psychological testing* (5e éd. rev.). New York : Harper Collins.

CRONBACH, L.J. et GLESER, G.C. (1965). *Psychological tests and personnel decisions* (2e éd. rev.). Urbana, IL : University of Illinois Press.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

CRONBACH, L.J. et MEEHL, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, *52* (4), 281-302.

DAWES, R.M. (1979). The robust beauty of improper linear models in decision-making. *American Psychologist*, *34*, 571-582.

DAWES, R.M. et CORRIGAN, B. (1974). Linear models in decision-making. *Psychological Bulletin*, *81*, 95-106.

DAWES, R.M., FAUST, D. et MEEHL, P.E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668-1674.

DEVELLIS, R.F. (1991). *Scale development*. Newbury Park, CA : Sage.

DÉVELOPPEMENT DES RESSOURCES HUMAINES DU CANADA (1993). *Classification nationale descriptive des professions*. Canada, Ottawa : Ministère des Approvisionnements et Services.

DÉVELOPPEMENT DES RESSOURCES HUMAINES DU CANADA (1995). *Logiciel de la classification nationale descriptive des professions*. Canada, Ottawa : Ministère des Approvisionnements et Services.

DIGMAN, J.M. (1990). Personality structure : Emergence of the five-factor model. *Annual Review of Psychology*, *41*, 417-440.

DREHER, G.F. et SACKETT, P.R. (1983). Commentary : A critical look at some common beliefs about assessment centers. *In* G.F. Dreher et P.R. Sackett (éd.). *Perspectives on employment staffing and selection* (p. 258-265). Homewood, IL : Irwin.

DUNNETTE, M.D. (1966). *Personnel selection and placement*. Wadsworth Publishing Company : Belmont, CA.

DUNNETTE, M.D. (1976). Aptitudes, abilities, and skills. *In* M.D. Dunnette (éd.). *Handbook of industrial and organizational psychology* (p. 473-520). Chicago : Rand McNally.

EEOC : voir Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, et Department of Justice.

EQUAL EMPLOYMENT OPPORTUNITY COMMISSION (EEOC), CIVIL SERVICE COMMISSION, DEPARTMENT OF LABOR, ET DEPARTMENT OF JUSTICE (1978). *Adoption by four agencies of uniform guidelines on employee selection procedures*. Federal Register, 43 (166), 38290-38315.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

EREZ, A., BLOOM, M.C. et WELLS, M.T. (1996). Using random rather than fixed effects models in meta-analysis : Implications for situational specificity and validity generalization. *Personnel Psychology, 49*, 275-307.

FITZ-ENZ, J. (1980). Quantifying the human resource function. *Personnel, 57*, mars-avril, 41-52.

FLANAGAN, J.C. (1954). The critical incident technique. *Psychological Bulletin, 51*, 327-355.

FLEISHMAN, E.A. et REILLY, M.E. (1992). *Handbook of human abilities*. Palo Alto, CA : Consulting Psychology Press.

FRIED, Y. et AGER, W. (1998). Meta-analysis : review, integration and recommendations for meta-analysts. *In* C.L. Cooper et I.T. Robertson (éd.). *International Review of Industrial and Organizational Psychology 1998, vol. 13* (p. 123-158). Chichester, Angleterre : John Wiley.

GATEWOOD, R.D. et FEILD, H.S. (1998). *Human resource selection* (4e éd. rev.). Fort Worth, TX : Harcourt Brace.

GAUGLER, B.B., ROSENTHAL, D.B., THORNTON III, G.C. et BENSON, C. (1987), Meta-analysis of assessment center validity. *Journal of Applied Psychology, 72*, 493-511.

GAVIN, A.T. (1977). *Guide to the development of written tests for selection and promotion : The content validity model* (Technical memorandum 77-6). Washington, D.C. : Personnel Research and Development Center, United States Civil Service Commission.

GHISELLI, E.E. (1966). *The validity of occupational aptitude tests*. New York : John Wiley and Sons.

GHISELLI, E.E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology, 26*, 461-477.

GOLDSTEIN, I.L., ZEDECK, S. et SCHNEIDER, B. (1993). An exploration of the job analysis-content validity process. *In* N. Schmitt, W.C. Borman and Associates (éd.). *Personnel selection in organization* (p. 3-34). San Francisco : Jossey-Bass.

GORDON, M.E. (1972). Three ways to effectively evaluate personnel programs. *Personnel Journal, 51*, juillet, 498-504.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

GOSSELIN, A. et MURPHY, K.R. (1994). L'échec de l'évaluation de la performance. *Gestion, Revue internationale de gestion*, *19* (3), 17-28.

GRANT, P.C. (1989). *Multiple use job descriptions : A guide to analysis, preparation, and applications for human resources managers*. Westport, CT : Quorum Books.

GUION, R.M. (1965). *Personnel testing*. New York : McGraw-Hill.

GUION, R.M. (1977). Content validity – The source of my discontent. *Applied Psychological Measurement*, *1* (1), 1-10.

GUION, R.M. (1991). Personnel assessment, selection, and placement. *In* M.D. Dunnette et L.M. Hough (éd.). *Handbook of industrial and organizational psychology* (2e éd. rev.), *vol. 2* (p. 327-397). Palo Alto, CA : Consulting Psychologists Press.

GUION, R.M. (1998). *Assessment, measurement, and prediction for personnel decisions*. Mahwah, NL : Lawrence Erlbaum Associates.

GUION, R.M. et CRANNY, C.J. (1982). A note on concurrent and predictive validity designs : A critical reanalysis. *Journal of Applied Psychology*, *67*, 239-244.

HAKEL, M.D. (1989). Merit-bases selection : Measuring the person for the job. *In* W.F. Cascio et D.H. Sweet (éd.). *Human resource planning employment and placement* (p. 135-158). Washington, D.C. : Bureau of National Affairs.

HARRIS, M.M. (1989). Reconsidering the employment interview : A review of recent literature and suggestions for future research. *Personnel Psychology*, *42*, 691-726.

HARTIGAN, J.A. et WIGDOR, A.K. (1989). *Fairness in employment testing*. Washington, D.C. : National Academy Press.

HARVEY, R.J. (1991). Job analysis. *In* M.D. Dunnette et L.M. Hough (éd.). *Handbook of industrial and organizational psychology* (2e éd. rev.), *vol. 2* (p. 71-163). Palo Alto, CA : Consulting Psychologists Press.

HELMSTADTER, C.C. (1964). *Principles of psychological measurement*. New York : Meredith Publishing.

HOFFMAN, C.C. (1995). Applying range restriction corrections using published norms : three case studies. *Personnel Psychology*, *48*, 913-923.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

HOFFMAN, C.C. et MCPHAIL, S.M. (1998). Exploring options for supporting test use in situations precluding local validation. *Personnel Psychology, 51*, 987-1003.

HOFFMAN, C.C., NATHAM, B.R. et HOLDEN, L.M. (1991). A comparison of validation criteria : Objective versus subjective performance measures and self versus supervisor. *Personnel Psychology, 44*, 601-619.

HOFFMAN, C.C. et THORNTON III, G.C. (1997). Examining selection utility where competing predictors differ in adverse impact. *Personnel Psychology, 50*, 455-470.

HOGAN, R.T. (1991). Personality and personality measurement. *In* M.D. Dunnette et L.M. Hough (éd.). *Handbook of industrial and organizational psychology* (2e éd. rev.), *vol. 2* (p. 873-919). Palo Alto, CA : Consulting Psychologists Press.

HOLLAND, J.L. (1973). *Making vocational choices : A theory of careers*. Englewood Cliffs, NJ : Prentice-Hall.

HUFFCUTT, A.I. et Arthur Jr., W. (1994). Hunter and Hunter (1984) revisited : Interview validity for entry-level jobs. *Journal of Applied Psychology, 79*, 184-190.

HUNT, M. (1997). *How science takes stock : The story of meta-analysis*. New York : Russell Sage Foundation.

HUNT, S.T. (1996). Generic work behavior : An investigation into the dimensions of entry-level, hourly job performance. *Personnel Psychology, 49*, 51-83.

HUNTER, J.E. (1981). *What is the validity of a content valid test ?* International Personnel Management Association, Minneapolis.

HUNTER, J.E. (1983). *Test validation for 12,000 jobs : An application of job classification and validity generalization analysis to the General Aptitude Test Battery* (USES test research report no. 45). Washington, D.C. : Division of Counselling and Test Development, Employment and Training Administration, U.S. Departement of Labor.

HUNTER, J.E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior, 29* (3), 340-362.

HUNTER, J.E. et HUNTER, R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96* (1), 72-98.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

HUNTER, J.E. et SCHMIDT, F.L. (1982). Fitting people to jobs : The impact of personnel selection on national productivity. *In* M.D. Dunnette et E.A. Fleishman (éd.). *Human performance and productivity, vol. 1* (p. 233-284). Hillsdale, NJ : Erlbaum.

HUNTER, J.E., SCHMIDT, F.L. et JUDIESCH, M.K. (1990). Individual differences in output variability as a function of job complexity. *Journal of Applied Psychology, 75*, 28-42.

HUSELID, M.A., JACKSON, S.E. et SCHULER, R.S. (1997). Technical and strategic human resource management effectiveness as determinants of firm performance. *Academy of Management Journal*, *40*, 171-188.

KELLOGG, C.E. et NORTON, N.W. (1978). *Manual – Revised Beta examination, second edition (Beta-II)*. New York : The Psychological Corporation.

KERLINGER, F.N. (1973). *Foundations of behavioral research* (2e éd. rev.). New York : Holt, Rinehart and Winston.

KERLINGER, F.N. (1986). *Foundations of behavioral research* (3e éd. rev.). New York : Holt, Rinehart and Winston.

KLEIMAN, L.S. et FALEY, R.H. (1985). The implications of professional and legal guidelines for court decisions involving criterion-related validity : A review and analysis. *Personnel Psychology, 38*, 803-833.

LANDY, F.J. et VASEY, J. (1991). Job analysis : The composition of SME samples. *Personnel Psychology, 44*, 27-50.

LAWSHE, C.H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563-575.

LATHAM, G.P. et WHYTE, G. (1994). The futility of utility analysis. *Personnel Journal*, *47*, 31-46.

LENT, R.H., AURBACH, H.A. et LEWIN, L.S. (1971). Predictors, criteria, and significant results. *Personnel Psychology, 24*, 519-533.

LEVINE, E.L., SPECTOR, P.E., MENON, S., NARAYANAN, L. et CANNON-BOWERS, J. (1996). Validity generalization for cognitive, psychomotor, and perceptual tests for craft jobs in the utility industry. *Human Performance*, *9*, 1-22.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

LOWENBERG, G, LOSCHENKOHL, G.H. et FAUST, B.D. (1985). Meta-analysis demonstrating validity generalization for managerial assessment center dimensions. 93rd Annual convention of American Psychological Association, Los Angeles, CA.

LUBINSKI, D. et DAWIS, R.V. (1992). Aptitudes, skills, and proficiencies. *In* M.D. Dunnette et L.M. Hough (éd.). *Handbook of industrial and organizational psychology* (2e éd. rev.), *vol. 3* (p. 1-59). Palo Alto, CA : Consulting Psychologists Press.

LUTHANS, F. et MARIS, T.L. (1979). Evaluating personnel programs through the reversal technique. *Personnel Journal*, *58*, October, 692-697.

MCCLELLAND, D.C. (1976). *A guide to job competence assessment*. Boston : McBer.

MCCORMICK, E.J. (1976). Job and task analysis. *In* M.D. Dunnette (éd.). *Handbook of industrial and organizational psychology* (p. 651-696). Chicago : Rand McNally.

MCCORMICK, E.J., MECHAM, R.C. et JEANNERET, P.R. (1989). *Technical manual for the Position Analysis Questionnaire (PAQ)* (2e éd. rev.). West Lafayette, IN : Purdue Research Foundation.

MCCORMICK, E.J. et TIFFIN, J. (1974). *Industrial psychology* (6e éd. rev.). Englewood Cliffs, NJ : Prentice-Hall.

MCDANIEL, M.A., WHETZEL, D.L., SCHMIDT, F.L. et MAURER, S.D. (1994). The validity of employment interview : A comprehensive review and meta-analysis. *Journal of Applied Psychology*, *79*, 599-616.

MCHENRY, J.J., HOUGH, L.M., TOQUAM, J.L., HANSON, M.A. et ASHWORTH, S. (1990). Project A validity results : The relationship between predictor and criterion domains. *Personnel Psychology*, *43*, 335-354.

MCMANUS, M.A. et KELLEY, M.L. (1999). Personality measures and biodata : Evidence regarding their incremental predictive value in the life insurance industry. *Personnel Psychology*, *52*, 137-148.

MAGNUSSON, D. (1966). *Test theory*. Reading, MA : Addison-Wesley.

MAIER, N.R.F. (1970). *La psychologie dans l'industrie*. Verviers, Belgique : Gérard et Co.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

MAILLET, L. (1993). *Psychologie et organisation* (2e éd. rev.). Laval, Québec, Canada : Éditions Agence D'Arc.

MARCHESE, M.C. et MUCHINSKY, P.M. (1993). The validity of the employment interview : A meta-analysis. *International Journal of Selection and Assessment, 1* (1), 18-26.

MAURER, T.J. et ALEXANDER, R.A. (1992). Methods of improving employment test critical scores derived by judging test content : A review and critique. *Personnel Psychology, 45*, 727-762.

MELANSON, D. et FONTAINE, L. (1993). *Vision ressources humaines*. Montréal : Sobeco Ernst et Young.

MERCER, M. (1989). *Turning your human resources department into a profit center.* New York : Amacom.

MINTZBERG, H. (1973). *The nature of managerial work*. New York : Harper and Row.

MINTZBERG, H. (1975). The manager's job : Folklore and fact. *Harvard Business Review,* juillet-août, 49-61.

MOUNT, M. et BARRICK, M. (1995). The big five personality dimensions : Implications for research and practice in human resources management. *Research in Personnel and Human Resources Management, 13*, 153-200.

MURPHY, J. et CLEVELAND, N. (1991). *Performance appraisal*. Needham, MA : Allyn and Bacon.

MURPHY, K.R. et DAVIDSHOFER, C.O. (1988). *Personnel testing*. Englewood Cliffs, NJ : Prentice-Hall.

MURPHY, K.R. et SHIARELLA, A.H. (1997). Implications of the multidimensional nature of job performance for the validity of selection tests : multivariate frameworks for studying test validity. *Personnel Psychology, 50*, 823-855.

MUSSIO, S.J. et SMITH, M.K. (1973). *Content validity : A procedural manual*. Minneapolis, MN : Minneapolis Civil Service Commission.

NATHAM, B.R. et ALEXANDER, R.A. (1988). A comparison of criteria for test validation : A meta-analytic investigation. *Personnel Psychology, 41*, 517-535.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

NUNNALLY, J.C. et BERNSTEIN, I.H. (1994). *Psychometric theory* (3e éd. rev.). New York : McGraw-Hill.

OUCHI, W. et WILKINS, A. (1985). Organizational culture. *Annual Review of Sociology, 11*, 457-483.

PETTERSEN, N. (1991). Selecting project managers : An integrated list of predictors. *Project Management Journal*, XXII (2), 21-26.

PETTERSEN, N. et JACOB, R. (1992). *Comprendre le comportement de l'individu au travail : un schéma d'intégration*. Laval, Québec, Canada : Éditions Agence D'Arc.

PLUMLEE, L.B. (1980). *A guide to the development of job knowledge and skill tests : A reference kit for measurement specialists, second edition* (Personnel research report 80-27). Washington, D.C. : Personnel Research and Development Center, United States Office of Personnel Management.

QUINONES, M.A., FORD, J.K. et TEACHOUT, M.S. (1995). The relationship between work experience and job performance : A conceptual and meta-analytic review. *Personnel Psychology, 48*, 887-910.

REE, M.J. et CARRETTA, T.R. (1998). General cognitive ability and occupational performance. *In* C.L. Cooper et I.T. Robertson (éd.). *International Review of Industrial and Organizational Psychology 1998, vol. 13* (p. 159-184). Chichester, Angleterre : John Wiley.

REE, M.J., CARRETTA, T.R., EARLES, J.A. et ALBERT, W. (1994). Sign changes when correcting for range restriction : A note on Pearson's and Lawley's selection formulas. *Journal of Applied Psychology, 79*, 298-301.

REE, M.J. et EARLES, J.A. (1991). Predicting training success : Not much more than *g*. *Personnel Psychology, 44*, 321-332.

REILLY, P.L. et CHAO, G.T. (1982). Validity and fairness of some alternative employee selection procedures. *Personnel Psychology, 35*, 1-62.

SARRAZIN, G. (1994). Épistémologie de la mesure : la théorie de la validité et ses conséquences. *L'orientation, 7* (1), 9-16.

SCHIPPMANN, J.S., PRIEN, E.P. et KATZ, J.A. (1991). Reliability and validity of in-basket performance measures. *Personnel Psychology, 43*, 837-859.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

SCHMITT, N., GOODING, R.Z., NOE, R.A. et KIRSCH, M. (1984). Meta-analyses of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407-422.

SCHMIDT, F.L. et HUNTER, J.E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology, 62*, 529-540.

SCHMIDT, F.L. et HUNTER, J.E. (1980). The future of criterion-related validity. *Personnel Psychology, 33*, 41-60.

SCHMIDT, F.L. et HUNTER, J.E. (1983). Individual differences in productivity : An empirical test of estimates derived from studies of selection procedure utility. *Journal of Applied Psychology, 68*, 407-414.

SCHMIDT, F.L. et HUNTER, J.E. (1998). The validity of selection methods in personnel psychology : Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

SCHMIDT, F.L., HUNTER, J.E. et PEARLMAN, K. (1982). Progress in validity generalization : Comments on Callender and Osburn and further developments. *Journal of Applied Psychology, 67*, 835-845.

SCHMIDT, F.L., HUNTER, J.E. et URRY, V.W. (1976). Statistical power in criterion related validation studies. *Journal of Applied Psychology, 61*, 473-485.

SCHMIDT, F.L., HUNTER, J.E., MCKENZIE, R.C. et MULDROW, T.W. (1979). Impact of valid selection procedures on work-force productivity. *Journal of Applied Psychology, 64*, (609-626.

SCHMIDT, F.L., HUNTER, J.E., OUTERBRIDGE, A.N. et GOFF, S. (1988). The economic impact of job selection methods on size, productivity, and payroll costs of the federal work force : An empirically based demonstration. *Personnel Psychology, 35*, 1-29.

SCHMIDT, F.L., HUNTER, J.E., OUTERBRIDGE, A.N. et TRATTNER, M.H. (1986). Joint relation of experience and ability with job performance : Test of three hypotheses. *Journal of Applied Psychology, 73*, 46-57.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

SCHMIDT, F.L., OCASIO, B.P., HILLERY, J.M. et HUNTER, J.E. (1985). Further within-setting empirical tests of the situational specificity hypothesis in personnel selection. *Personnel Psychology, 38*, 509-524.

SCHMIDT, F.L., PEARLMAN, K., HUNTER, J.E. et HIRSH, H.R. (1985). Forty questions about validity generalization and meta-analysis. *Personnel Psychology, 38*, 697-789.

SCHMITT, N. et LANDY, F.J. (1993). The concept of validity. *In* N. Schmitt, W.C. Borman and Associates (éd.). *Personnel selection in organization* (p. 275-309). San Francisco : Jossey-Bass.

SCHMITT, N. et CHAN, D. (1998). *Personnel selection*. Thousand Oaks, CA : Sage.

SCHNEIDER, B. et SCHMITT, N. (1986). *Staffing organizations* (2e éd. rev.). Glenview, IL : Scott, Foresman and Company.

SCHULER, H. et GULDIN, A. (1991). Methodological issues in personnel selection research. *In* C.L. Cooper et I.T. Robertson (éd.). *International Review of Industrial and Organizational Psychology 1991, vol. 6* (p. 213-264). Chichester, Angleterre : John Wiley.

SCIENCE RESEARCH ASSOCIATES (1973). *Validation : Procedures and results – Part I : Procedures for test validation studies*. Chicago, IL : Science Research Associates Inc.

SCP : voir sous Société canadienne de psychologie.

SELLTIZ, C., WRIGHTSMAN, L.S. et COOK, S.W. (1976). *Research methods in social relations* (3e éd. rev.). New York : Holt, Rinehart et Winston.

SIOP : voir sous Society for Industrial and Organizational Psychology.

SMITH, M., GREGG, M. et ANDREWS, D. (1989). *Savoir recruter*. Paris : Eyrolles (1990).

SOCIÉTÉ CANADIENNE DE PSYCHOLOGIE (1987). *Lignes directrices pour les tests psychologiques et pédagogiques*.

SOCIETY FOR INDUSTRIAL AND ORGANIZATIONAL PSYCHOLOGY (SIOP) (1987). *Principles for validation and use of personnel selection procedures* (3e éd. rev.). College Park, MD.

SLINVINSKI, L.W. et MILES, J. (1996). *Profil global de compétence : un modèle*. Ottawa : Centre de psychologie du personnel.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

SPECTOR, P.E., BRANNICK, M.T. et COOVERT, M.D. (1988). Job analysis. *In* C.L. Cooper et I.T. Robertson (éd.). *International Review of Industrial and Organizational Psychology 1988, vol. 4* (p. 281-328). Chichester, Angleterre : John Wiley.

SPENCER, L.M. et SPENCER, S.M. (1993). *Competence at work*. New York : John Wiley.

STANDARDS… : voir sous American Educational Research Association *et al.*

SUSSMANN, M. et ROBERTSON, D.U. (1986). The validity of validity : An analysis of validation study designs. *Journal of Applied Psychology, 71*, 461-468.

TAYLOR, H.C. et RUSSEL, J.T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection : Discussion and tables. *Journal of Applied Psychology, 23*, 565-578.

TETT, R.P., JACKSON, D.N. et ROTHSTEIN, M. (1991). Personality measures as predictor of job performance : A meta-analytic review. *Personnel Psychology, 44*, 703-742.

TETT, R.P., MEYER, J.P. et ROESE, N.J. (1994). Applications of meta-analysis : 1987-1992. *In* C.L. Cooper et I.T. Robertson (éd.). *International Review of Industrial and Organizational Psychology 1994, vol. 9* (p. 71-112). Chichester, Angleterre : John Wiley.

THOMPSON, D.E. et THOMPSON, T.A. (1982). Court standards for job analysis in test validation. *Personnel Psychology, 35*, 865-874.

THORNDIKE, R.L. (1982). *Applied psychometrics*. Boston, MA : Houghton Mifflin.

THORNDIKE, R.M., CUNNINGHAM, G.K., THORNDIKE, R.L. et HAGEN, E. (1991). *Measurement and evaluation in psychology and education* (5e éd. rev.). New York : Macmillan.

THORNTON, G.C. (1992). *Assessment centers in human ressources management*. Reading, MA : Addison-Wesley.

TSUI, A.S. (1984). Personnel department effectiveness : a tripartite approach. *Industrial Relations, 23* (2), 184-197.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

TSUI, A.S. (1987). Defining the activities and effectiveness of the human resource department : A multiple constituency approach. *Human Resource Management, 26* (1), 35-69.

TZINER, A., JEANRIE, C. et CUSSON, S. (1993). *La sélection du personnel – Concepts et application.* Laval, Québec, Canada : Éditions Agence D'Arc.

UNITED STATES DEPARTMENT OF LABOR (1970). *Manual for the USES General Aptitude Test Beattery. Section III : Development.* Washington, D.C. : U.S. Department of Labor.

VÉZINA, C. (1979). *Les clauses d'ancienneté et d'arbitrage de griefs*. Ottawa : Éditions de l'Université d'Ottawa.

WEISNER, W.H. et CRONSHAW, S.F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology, 61*, 275-290.

WERNIMONT, P.F. et CAMPBELL, J.P. (1968). Signs, samples, and criteria. *Journal of Applied Psychology, 52*, 372-376.

WIGDOR, A.K. et GARNER, W.R. (1982). *Ability testing : Uses, consequences, and controversies*. Washington, D.C. : National Academy Press.

WRIGHT, P.M., LICHTENFELS, P.A. et PURSELL, E.D. (1989). The structured interview : Additional studies and meta-analysis. *British of Occupational Psychology, 62*, 191-199.

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*,
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

INDEX

© 2000 – Presses de l'Université du Québec

Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure*, Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés

© 2000 – Presses de l'Université du Québec
Édifice Le Delta I, 2875, boul. Laurier, bureau 450, Québec, Québec G1V 2M2 • Tél. : (418) 657-4399 – www.puq.ca
Tiré : *Évaluation du potentiel humain dans les organisations : élaboration et validation d'instruments de mesure,*
Normand Pettersen, ISBN 2-7605-1051-4 • D1051N

Tous droits de reproduction, de traduction ou d'adaptation réservés