COMPTE-RENDU CRITIQUE

Laveault, D. et Grégoire, J. (2023). *Introduction aux théories des tests en psychologie et en sciences de l'éducation* (4e éd.). De Boeck Supérieur.

Karine Paquette-Côté
paquette-cote.karine@courrier.uqam.ca
Université du Québec à Montréal

Il n'est pas aisé de faire la recension de Laveault et Grégoire. D'abord, comme cet ouvrage en est à sa quatrième édition, c'est une œuvre de carrière universitaire que bien d'autres praticiens et chercheurs ont eu l'occasion de lire, d'utiliser et de commenter. C'est également un ouvrage parmi les rares en français qui est consacré à un sujet aussi complexe et aussi important que le développement des tests et des questionnaires, une pratique courante et pourtant trop souvent négligée par apparente simplicité et commun usage. L'accessibilité du savoir en psychométrie et en édumétrie étant d'une grande nécessité, les livres d'introduction dans le domaine devraient simplifier les notions spécialisées encore trop souvent réservées à une poignée d'experts. Il est évident que c'est ce à quoi se sont consacrés Laveault et Grégoire, mais force est d'admettre que le fossé demeure encore trop grand entre experts et non-initiés. Cet ouvrage n'est pas une introduction à proprement parler. S'il l'a déjà été, les nombreux retraits et remplacements de contenu au fil des éditions, pour lui permettre de conserver une ampleur d'un peu moins de 400 pages, auront possiblement contraint les auteurs à retirer plusieurs notions de base pour couvrir davantage de spécialités.

La première phrase de l'avant-propos énonce l'essence de l'ouvrage et de chacune de ses éditions : « refléter les progrès accomplis dans le domaine de la mesure et de l'évaluation tout en réaffirmant les fondements théoriques en psychométrie et en édumétrie qui ont particulièrement bien

traversé les années ». Les spécialistes s'accorderont sans doute pour dire que les modèles classiques ont particulièrement bien traversés les années, si bien qu'il est difficile de s'en distancer!

L'ouvrage est divisé en sept chapitres couvrant l'un après l'autre les thèmes de la construction d'un test, les scores et leur distribution, la fidélité des résultats, la validité des résultats, l'analyse des items, la transformation et l'interprétation des scores et, enfin, les modèles de réponse à l'item et les nouvelles technologies. Une annexe présente des tables statistiques et un glossaire des termes techniques et de traductions en anglais et en français. Ce dernier sera très apprécié des lecteurs de littérature scientifique anglophone spécialisée dans le domaine, de même que des autodidactes. En matériel complémentaire, des tests en ligne fournissent des exemples d'items variés et très bien construits (à reproduire et parfois, à éviter) et constituent une forme ludique de révision du contenu de chaque chapitre. Des items consistant à trouver les valeurs manquantes d'une formule pourraient toutefois manquer de pertinence. L'accès à ces tests en ligne est facilité par la présentation d'un lien court et d'un code QR à la fin de chaque chapitre. Toutefois, bien que le texte fasse référence à un document en ligne intitulé *Notions d'inférence statistique*, aucun lien n'est fourni dans l'ouvrage ni sur la page descriptive du livre chez l'éditeur De Boeck Supérieur. En ce qui concerne l'organisation du contenu, la lecture serait facilitée par l'ajout d'une introduction mettant en contexte la psychométrie et l'édumétrie en psychologie et en éducation et reprenant, au chapitre 1, ses aspects les plus généraux. Aussi, la lecture du deuxième chapitre pourrait avantageusement précéder celle du premier qui, lui, entre rapidement dans le vif du sujet.

Au chapitre 1, les auteurs prennent soin de décrire et d'illustrer les étapes de la construction d'un instrument de mesure au moyen d'exemples concrets, ce qui sera d'une aide précieuse lors d'usages professionnels et pratiques. Ce chapitre constitue une nouveauté par rapport aux éditions précédentes. L'approche n'est pas sans rappeler celle qu'ont employée Irwing et Hughes (2018) dans leur chapitre consacré au développement d'un test, lequel introduit le *Wiley Handbook of Psychometric Testing* qui rassemble en deux volumes des textes de référence en anglais rédigés par des spécialistes renommés qui ont uni leur plume autour d'un objectif commun : rendre accessibles les connaissances de pointe du domaine de la psychométrie. Alors que Irwing et Hughes couvrent le développement du test en 10 grandes étapes génériques dont la description introduit à tout

contexte de développement et d'utilisation d'un test de façon multidisciplinaire, Laveault et Grégoire se concentrent sur les cinq premières étapes de Downing (2006) et sur leur application concrète au domaine de l'éducation. Ce choix peut d'ailleurs amener les lecteurs du domaine de la psychologie à se questionner sur la pertinence de l'ouvrage pour eux puisqu'il leur faudra lire 75 pages avant de trouver les premiers exemples appliqués spécifiquement à leur domaine. Certains exemples pourraient d'ailleurs être difficiles à comprendre en première lecture parce qu'ils nécessitent très tôt la compréhension de notions complexes expliquées bien plus loin dans l'ouvrage. C'est le cas, par exemple, d'une interprétation bien spécifique et restreinte de la notion de validité (plus précisément, la validité critériée) qui est utilisée au chapitre 1 et dont la compréhension requiert la lecture hâtive de son explication au chapitre 5. Malgré ces petits défauts de structure, ce premier chapitre fait la démonstration de la complexité de la pratique évaluative, une pratique exigée de tout enseignant et qui, pourtant, n'occupe pas la place et ne retient pas toute l'attention qu'elle devrait dans leur cursus de formation. Plusieurs phrases de ce chapitre sonnent comme de la musique aux oreilles des passionnés : « […] il n'y a pas de bon format d'item dans l'absolu. Un format est bon s'il est adéquat au but et à la situation d'évaluation » (Laveault & Grégoire, 2023, p. 12) ou encore « […] la validité n'est jamais une qualité acquise une fois pour toutes. Chaque nouvelle inférence qu'un praticien veut réaliser à partir des résultats doit faire l'objet d'une validation spécifique » (Laveault & Grégoire, 2023, p. 16). Les auteurs présentent également des notions pédagogiques cohérentes telles que l'alignement pédagogique et l'évaluation authentique, ainsi que des plus populaires et toujours fortement utilisées comme les taxonomies d'objectifs. Les pédagogues et les technopédagogues avertis y trouveront des sources d'inspiration. Le personnel enseignant débutant pourrait toutefois manquer d'explications pour s'approprier les pratiques décrites, et ce, même si des exemples sont destinés à faciliter la compréhension des notions présentées. Les auteurs démontrent de mille et une façons que l'évaluation en éducation repose essentiellement sur la justification des choix par les praticiens davantage que sur un répertoire de bonnes ou de mauvaises pratiques. Malgré cela, des conseils de construction d'items sont énoncés, souvent en amorce de paragraphe, et le regroupement de ces conseils dans un tableau synthèse rendrait probablement l'ouvrage plus accessible et plus attirant.

Le chapitre 2 consiste en une révision des notions essentielles de la statistique descriptive nécessaires à la compréhension des notions plus avancées abordées ensuite. Les auteurs suggèrent aux personnes les plus averties d'en sauter la lecture si les notions abordées leurs sont familières. Ce chapitre fournit d'emblée des exemples de mesure issues du quotidien qui s'appliquent à la fois au domaine de la psychologie et à celui de l'éducation. La matière est très bien vulgarisée. Le chapitre fait un tour d'horizon permettant à quiconque de comprendre les notions essentielles de la statistique descriptive. Toute personne intéressée aux raisonnements mathématiques qui sous-tendent certaines méthodes et certaines techniques d'analyse statistique sera bien servie. Pour les autres, un survol rapide de certains passages ne nuira pas à la compréhension globale nécessaire à l'application de ces techniques.

Le chapitre 3 sur la fidélité des résultats présente des explications et des illustrations d'une rare clarté au sujet de la composition de la variance et à propos de la distinction entre fidélité et validité dans la théorie classique des scores. Toutefois, il est regrettable qu'un seul coefficient de fidélité soit abordé, tant dans la définition théorique que la définition opérationnelle qu'en font les auteurs (Laveault & Grégoire, 2023, p. 139-140). Le concept de fidélité ne se résume pas au coefficient de « corrélation entre les scores observés à deux formes parallèles » d'un test. Comme le soulignent Revelle et Condon (2018, p. 715-716), « les estimations de fidélité peuvent être basées sur des variations de l›ensemble du test, des variations dans le temps, des variations entre les éléments d›un test et la variabilité associée à la personne qui fait passer le test. Chacune de ces options a une signification différente et parfois un certain nombre d'estimations différentes » [notre traduction]. Ces auteurs présentent d'ailleurs une organisation des différentes interprétations de la fidélité et des coefficients qui y sont associés, dont certains sont expliqués par Laveault et Grégoire. Les coefficients de corrélation de cohérence interne par la méthode Spearman-Brown et celle de Rulon-Guttman sont cités pour leur utilisation pour vérifier l'équivalence des résultats de deux formes différentes d'un même test ou à partir de différentes bissections d'un test. Considérant sa popularité, une grande importance est accordée à l'alpha de Cronbach ou à des méthodes apparentées comme le Kuder-Richardson-20. Les auteurs prennent soin d'énoncer les limites de l'utilisation du coefficient alpha de Cronbach, qui est fortement critiquée et souvent inappropriée (Béland et al., 2017; Bourque et al., 2019; Laveault, 2012). Le coefficient oméga de

McDonald est présenté comme une option plus adéquate en éducation et en psychologie. La prochaine édition fera peut-être mention de plus d'un coefficient omega et référera peut-être à la contribution distincte d'un facteur général et de facteurs spécifiques dans le calcul des coefficients omega (Dueber & Toland, 2023). Les corrélations intra-classes issues de la théorie de la généralisabilité sont ensuite présentées pour permettre l'estimation des accords interjuges. Laveault et Grégoire ne couvrent pas les six coefficients Lambda de Guttman (1945) et la plus grande limite inférieure (en anglais *greatest lower bound*), dont le lecteur intéressé pourra approfondir l'exploration en consultant Revelle (2001/2023). Le chapitre 3 de Laveault et Grégoire se poursuit avec la description des limites à l'interprétation d'un coefficient de fidélité : la difficulté d'un test, l'étendue des différences individuelles, la limite de temps d'un test chronométré et la longueur du test. Suit la description d'une procédure de construction d'un intervalle de confiance autour de l'estimation du score vrai qui, bien que rigoureusement décrite, semble déconnectée de la pratique. Il en est de même pour l'erreur type d'estimation. L'erreur type de la différence est bien décrite avec un exemple concret. La section sur le modèle binomial de l'erreur est essentiellement une démonstration statistique visant à permettre de contourner les limites de la théorie classique et, ainsi, tenter de permettre le calcul de l'erreur type de mesure pour différents niveaux de scores. À l'issue de sa lecture, une question s'impose : à quel public est destinée cette section ? La partie suivante contraste nettement puisqu'elle présente des cas concrets d'applications de l'étude de la généralisabilité. En effet, à partir de cette section, le ton change radicalement, tant dans la forme que dans le contenu : les manipulations de formules laissent place aux concepts entourant l'importante notion de variance et de ses composantes. Celles-ci dépassent l'instrument de mesure pour inclure les personnes impliquées dans la situation de mesure, de même que différentes caractéristiques de celle-ci, chacune constituant une source d'erreur distincte responsable d'une partie de la variance totale des scores. Les auteurs situent ensuite l'analyse de cette part d'erreur en fonction des objectifs poursuivis par la mesure, amenant à distinguer l'erreur relative et l'erreur absolue, et, ainsi, introduire l'importante notion de seuil. La simulation d'une étude de généralisabilité permet de démontrer, de façon classique, une situation de mesure où sont considérées les facettes des participants, de la tâche d'évaluation et des juges, ainsi que leurs interactions. Cette démonstration occupe beaucoup d'espace, mais l'attention accordée à ces notions est

justifiée. En effet, la théorie de la généralisabilité a permis de mettre en lumière toute la complexité des situations de mesure et elle mérite qu'on s'y attarde. Mais, bien qu'extrêmement rationnelles, ces situations illustrant des applications de la théorie classique et de la théorie de la généralisabilité « souffrent cependant d'un manque de conformité à la réalité. Soyons honnêtes : quand peut-on réellement observer deux tests parallèles ? En outre, jusqu'à quel point peut-on estimer avec une précision suffisante un coefficient de généralisabilité d'un plan d'observation à cinq ou six facettes croisées ? » (Bertrand & Blais, 2004, p. 108).

Si le chapitre sur la fidélité va parfois un peu loin dans la démonstration, le contraire pourrait être dit au sujet du chapitre 4, sur la validité des résultats. Le chapitre débute par une mise en contexte historique de l'évolution des concepts de validité et de validation pour enchaîner sur la description des catégories de preuves de validité à recueillir et à évaluer pour soutenir et démontrer la validité de l'interprétation des scores au test. On ne saurait le répéter suffisamment : « ce ne sont pas les tests qui sont valides, ni même les scores à ces tests, mais les interprétations et les décisions basées sur les scores » (Laveault & Grégoire, 2023, p. 199). Les auteurs abordent les preuves reposant sur le contenu, les processus de réponse, la structure interne du test, les relatons avec d'autres variables et les conséquences de l'utilisation des tests.

Les preuves de validité reposant sur le contenu, telles que décrites par Laveault et Grégoire, sont centrées sur le jugement d'experts et sur la démonstration de l'analyse quantitative de l'accord interjuges. La discussion concernant la validité d'un contenu pour mesurer un concept dans un contexte donné et dans la poursuite de certains objectifs identifiés est « dans une certaine mesure relative aux standards humains et aux cognitions » [notre traduction] (Markus & Borsboom, 2013, p. 281). En ce sens, elle est essentiellement qualitative. Le chapitre aurait gagné à mettre davantage l'accent sur les processus humains de recherche, d'analyse et d'argumentation par les experts qui soutiennent les liens causaux présumés unir entre eux les items et les autres composantes de la situation de mesure ainsi que d'autres construits pouvant leur être liés dans une théorie prédictive. La description de l'analyse factorielle et de l'origine historique de son développement est d'une rare clarté et constitue un exercice de synthèse étonnant. On peut tout de même s'interroger sur le fait que la rotation orthogonale soit présentée d'emblée, alors que la grande majorité des phénomènes étudiés en éducation et en psychologie, domaines d'application de

l'ouvrage, nécessite de tenir compte de leur corrélation, ce que permettent de nombreuses méthodes, dont l'ESEM. Les modèles d'équations structurelles, ou modèles structuraux d'équations, tels que nommés par Laveault et Grégoire, sont d'ailleurs très sommairement présentés, à la manière d'une introduction. La section sur la validité différentielle amène ensuite à questionner la validité des inférences produites à partir des résultats au test lorsque soumis à différents groupes de personnes.

Présentée au chapitre 5, l'analyse des items est centrée sur les indices de difficulté, de discrimination, de fidélité et de validité. Les auteurs décrivent avec finesse ce qui se cache derrière les résultats au test et qui reste obscur pour plusieurs praticiens. La présentation nous permet de constater que les résultats devraient toujours être interprétés en fonction des objectifs poursuivis par l'évaluation. Encore une fois, Laveault et Grégoire ont su bien vulgariser les concepts et les indices de l'analyse d'items. Considérant qu'ils ont dû faire preuve de parcimonie lors des passages d'une édition à l'autre, certains choix de contenus sont discutables, comme la section sur l'équivalence des items qui se termine sur l'incapacité de la méthode décrite à démontrer l'appartenance de différents items à un même domaine. La section sur l'analyse d'items uniques montre toutefois très pertinemment comment un item unique peut constituer une mesure aussi valide que peut l'être un questionnaire à multiples items. L'exemple de l'analyse d'items à l'aide d'IBM *SPSS Statistics* est possiblement l'extrait le plus utile et le plus recherché du public cible de l'ouvrage. En effet, un livre d'introduction à la psychométrie et à l'édumétrie devrait avant tout se consacrer à répondre aux questions : Que faire ? Pourquoi le faire ? Et, comment le faire ? II est tout de même regrettable que l'exemple donné comporte encore des items dichotomiques. En effet, bien que ces derniers soient souvent utilisés en éducation, les exemples d'analyse d'items de format polytomiques manquent toujours aux ouvrages méthodologiques, surtout en français, alors qu'ils sont fréquemment utilisés en psychométrie. À la suite de cet exemple, les auteurs présentent l'importante distinction entre la notion classique de biais et celle aujourd'hui largement acceptée de fonctionnement différentiel d'item. Le chapitre 5 se conclut sur la toute aussi pertinente mention selon laquelle l'analyse d'item doit être choisie en fonction des buts poursuivis. On ne saurait trop le répéter.

Le chapitre 6 porte sur la transformation et sur l'interprétation des scores. C'est à ce chapitre que le lecteur pourra enfin commencer à y voir clair : « Ce n'est pas parce que […] deux personnes ont le même score total

qu'elles ont nécessairement répondu de la même manière aux questions ou ont réussi aux mêmes items » (Laveault & Grégoire, 2023, p. 281). Ce chapitre aborde les patrons de réponses, les normes et leur relativité, les requis, les étapes et les pièges de la définition d'une population de référence et de l'échantillonnage, la transformation des scores bruts sur une échelle exprimant une norme. La section suivante aborde la mise en équivalence des scores à différents tests mesurant une même réalité, ce qui peut poser des défis d'acceptabilité sociale en contexte d'enseignement. Le chapitre se termine sur l'important concept de seuil de performance. Laveault et Grégoire présentent sommairement six méthodes fréquemment utilisées pour déterminer un seuil : quatre reposent sur le contenu du test et deux sur la performance des personnes. Encore une fois, le jugement d'experts est sollicité pour déterminer les choix de réponse pouvant être repérés comme incorrects par une personne détenant une habileté minimale, pour ensuite calculer une probabilité de réussite, estimer directement une probabilité de réussite, ou encore, juger de l'importance de la réussite des items en fonction du programme d'apprentissage. La validité diagnostique présentée en fin de chapitre permet d'évaluer la capacité d'un test à détecter la présence ou la probabilité d'apparition d'un trouble chez la personne.

L'ouvrage se termine par la théorie de la réponse à l'item. Ce septième et dernier chapitre contraste avec les autres, tant par son contenu que par sa présentation. Il commence par la présentation du concept de trait latent, puis de la courbe caractéristique de l'item, des paramètres de difficulté, de discrimination et de pseudo-chance et des exigences ou des postulats sur lesquels reposent les modèles de réponse à l'item. Le chapitre aborde ensuite la fonction d'information de l'item et du test. Les auteurs rappellent une fois de plus, et avec raison, que le but recherché par l'évaluation doit guider son analyse et son interprétation. Le chapitre se poursuit autour de l'évaluation de la dimensionnalité nécessaire pour s'assurer de respecter le postulat d'unidimensionnalité des modèles de réponse à l'item. Vient ensuite la description de l'analyse du fonctionnement différentiel d'item à partir de l'observation des courbes caractéristiques d'item pour différents sous-groupes de la population. Une grande partie du chapitre est ensuite réservée à l'usage des nouvelles technologies pour élaborer et administrer les tests. Cette partie n'est pas sans rappeler les ouvrages sur le même thème dirigés par Jean-Guy Blais et publiés en 2008 et en 2011.

En somme, la quatrième édition de l'ouvrage de Laveault et Grégoire couvre la vaste étendue des méthodes de la théorie classique des tests et introduit la théorie de la réponse à l'item. L'ouvrage est rédigé de façon à expliquer des concepts fondamentaux que les auteurs vulgarisent avec habileté et à amener la personne qui en fait la lecture à questionner les techniques d'analyse et à explorer ce qui les sous-tend. Cette exploration va parfois trop en profondeur pour un livre d'introduction. Il est surprenant de trouver certaines longues démonstrations statistiques pour des indices que la communauté scientifique n'utilise pratiquement plus alors que d'autres techniques en plein essor manquent clairement d'espace. La structure de l'ouvrage et la fluidité de lecture souffrent d'ailleurs des choix éditoriaux qu'ont dû faire les auteurs au fil des éditions. Il faut reconnaître que cet ouvrage est le fruit d'un travail colossal, une œuvre de carrière, une référence francophone à conserver à côté d'ouvrages spécialisés du domaine.

Comme l'écrivaient Bertrand et Blais (2004, p. 181) dans leur ouvrage sur les modèles de mesure : « Le travail du praticien et du développeur qui désire appliquer les modèles et s'assurer qu'il le fait adéquatement se trouve compliqué pour la simple et bonne raison qu'il est difficile à l'heure actuelle de choisir entre plusieurs propositions qui ont les mêmes prétentions, mais pour lesquelles le consensus théorique et empirique ne se réalise point ». Il faut reconnaître aux auteurs une grande réussite : amener les praticiens et les chercheurs à questionner sans relâche le contexte de la mesure et les objectifs qu'elle poursuit. Il n'y aura jamais trop d'ouvrages ou de carrières consacrés à cet objectif!

LISTE DES RÉFÉRENCES

Béland, S, Cousineau, D. & Loye, N. (2017). Utiliser le coefficient Omega de McDonald à la place de Cronbach. *Revue des sciences de l'éducation de McGill*, *52*(3), 791-804.

Bertrand, R. & Blais, J.-G. (2004). *Modèles de mesure. L'apport de la théorie des réponses* aux items. Presses de l'Université du Québec.

Blais, J.-G. (2008). Évaluation des apprentissages et technologies de l'information et de la communication. Enjeux, applications et modèles de mesure. Presses de l'Université Laval.

Blais, J.-G. & Gilles, J.-L. (2011). Évaluation des apprentissages et technologies de l'information et de la communication. Le futur est à notre porte. Presses de l'Université Laval.

Bourque, J., Doucet, D. LeBlanc, J., Dupuis. J. & Nadeau, J. (2019). L'alpha de Cronbach est l'un des pires estimateurs de la consistance interne : une étude de simulation. *Revue des sciences de l'éducation*, *45*(2), 78-99.

Downing, S. M. (2006). Twelve steps for effective test development. Dans S. M. Downing & T. M. Haladyna (dir.). *Handbook of test development*. Routledge.

Dueber, D. M. & Toland, M. D. (2023). A bifactor approach to subscore analysis. *Psychological Methods*. http://doi.org/10.1037/met0000459

Guttman, L. (1945). A basis for analysing test-retest reliability. *Psychometrika*, *2*(1), 41-54.

Irwing, P. & Hughes, D. J. (2018). Test Development. Dans P. Irwing, T. Booth & D. J. Hughes (dir.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (709-750). Wiley.

Laveault, D. (2012). Soixante ans de bons et mauvais usages du alpha de Cronbach. *Mesure et évaluation en éducation*, *35*(2), 1-7.

Markus, K. A. & Borsboom, D. (2013). Frontiers of test validity theory: Measurement, causation, and meaning. Routledge.

Revelle, W. (2001/2023). The personality project (chap. 7). Classical test theory and the measurement of reliability. https://www.personality-project.org/r/book/Chapter7.pdf

Revelle, W. & Condon, D. M. (2018). Reliability. Dans P. Irwing, T. Booth & D. J. Hughes (dir.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (709-750). Wiley.