Mesure et évaluation en éducation

La validité psychométrique : un regard global sur le concept centenaire, sa genèse, ses avatars

Nathalie André, Nathalie Loye and Louis Laurencelle

Volume 37, Number 3, 2015

URI: https://id.erudit.org/iderudit/1036330ar
DOI: https://doi.org/10.7202/1036330ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (print)
2368-2000 (digital)

Explore this journal

Cite this article

André, N., Loye, N. & Laurencelle, L. (2015). La validité psychométrique : un regard global sur le concept centenaire, sa genèse, ses avatars. *Mesure et évaluation en éducation*, *37*(3), 125–148. https://doi.org/10.7202/1036330ar

Article abstract

Since Alfred Binet, who, without mentioning validity explicitly, presented a pragmatic, utilitarian and empirical vision of the relevance of tests, the concept of validity of psychological tests has greatly evolved. In a historical perspective on the concept of psychometric validity, this paper aims to explore various facets in order to identify their wide definitional orientations, without ignoring the operational procedures on which they are based.

La validité psychométrique : un regard global sur le concept centenaire, sa genèse, ses avatars

Nathalie André
Université de Poitiers

Nathalie Loye
Université de Montréal

Louis Laurencelle
Université du Québec à Trois-Rivières

Mots clés : validité, validation, test psychométrique, construit, mesure

Depuis Alfred Binet, qui, sans parler de validité, présentait toutefois une conception pragmatique, utilitaire et empirique de la pertinence d'un test, le concept de validité est né et a beaucoup évolué. À partir d'une perspective historique du concept de validité psychométrique, cet article vise à en explorer de manière critique quelques facettes afin de dégager les différentes orientations définitionnelles, sans perdre de vue les démarches d'opérationnalisation qu'on leur associe.

Key words: validity, validation, psychometric test, construct, measurement

Since Alfred Binet, who, without mentioning validity explicitly, presented a pragmatic, utilitarian and empirical vision of the relevance of tests, the concept of validity of psychological tests has greatly evolved. In a historical perspective on the concept of psychometric validity, this paper aims to explore various facets in order to identify their wide definitional orientations, without ignoring the operational procedures on which they are based.

Palavras-chave: validade, validação, teste psicométrico, constructo, medição

Depois de Alfred Binet, o qual, sem falar da validade, apresentou uma conceção pragmática, utilitária e emprírica da pertinência de um teste, o conceito de validade nasceu e evoluiu significativamente. A partir de uma perspetiva histórica do conceito de validade psicométrica, este artigo visa explorar criticamente várias facetas para identificar as diferentes orientações definicionais, sem perder de vista os procedimentos de operacionalização nos quais se baseiam.

Note des auteurs : La correspondance liée à cet article peut être adressée à : Nathalie André, Université de Poitiers, [nathalie.andre@univ-poitiers.fr] ; Nathalie Loye, Université de Montréal, [nathalie.loye@umontreal.ca] ; Louis Laurencelle, UQTR, [louis.laurencelle@gmail.com].

Préambule

Le domaine des applications de la psychométrie et de ses concepts déborde largement celui dont il est issu, à savoir celui des échelles psychologiques. C'est particulièrement le cas en éducation, un secteur dans lequel pratiquement toutes les mesures sont obtenues à partir de réponses à un questionnaire. Or, le test de Binet-Simon (1905) présentait, comme le font aujourd'hui encore les tests de quotient intellectuel (QI), des tâches à réaliser, des épreuves chronométrées, etc. Cette même psychométrie, avec ses concepts de fidélité, de validité et de normes de classement, est aujourd'hui employée en médecine, en kinésiologie et en génie, pour ne citer que ces disciplines, là où l'objet de référence est généralement plus concret que l'introversion/extraversion, la motivation, l'habileté visuospatiale, voire l'intelligence.

Ainsi, en matière de validité psychométrique, il serait incorrect de se replier sur le seul secteur des qualités culturelles, interactionnelles ou strictement interprétatives des personnes. Il sied plutôt que les définitions et les arguments sur la valeur des tests englobent divers domaines de mesure des qualités vivantes de la personne.

L'essai de réflexion développé dans cet article repose sur une perspective historique du concept de validité psychométrique. Il vise à explorer de manière critique quelques facettes du concept afin d'en montrer l'évolution et les univers de réalisation ; de faire état des doctrines unificatrices proposées et de leur pertinence ; et, enfin et surtout, de faire voir la richesse de ce concept qui, en fait, est polysémique parce qu'entendu de multiples manières et rapportable aux différents contextes dans lesquels les tests sont appliqués.

La première partie du texte entame cette perspective historique en prenant comme fil conducteur la mesure de l'intelligence. Les premiers questionnements sont articulés autour du concept de mesure, de l'évolution des modélisations des données et de la conception des instruments. La deuxième partie vise à explorer diverses facettes du concept de validité. La dernière partie permet de porter un regard critique à partir de quelques études de validité inscrites dans le domaine de l'éducation. La conclusion porte sur l'état actuel de la réflexion engagée par cet article.

Les premiers questionnements issus du domaine de la psychologie

Autour du concept de mesure

Dans ses travaux destinés à cerner les aptitudes des enfants et à mesurer leur intelligence, Alfred Binet s'est largement inspiré de ceux effectués par William James sur les émotions. À ce propos, il écrivait :

> On a pu critiquer cette théorie, mais il a bien fallu reconnaître qu'elle est claire ; avec James, l'émotion cesse d'être un mot, une conception abstraite, c'est quelque chose d'intelligible et pour ainsi dire de tangible ; il n'a pas cherché à comprendre son rôle dans le mécanisme de la pensée, mais à saisir en quoi elle consiste, de quelle matière elle est faite ; il n'a pas présenté une théorie dynamique de l'émotion, mais plutôt une théorie statique, une définition, une analyse, un inventaire (Binet, 1910, p. 5-6).

Peut-on affirmer pour autant que la mesure des émotions, grâce à une interprétation subjective de phénomènes physiologiques éventuellement observables, ou de l'intelligence, par la mesure de comportements verbaux ou moteurs à composantes mentales complexes, en fait des concepts psychométriques, c'est-à-dire des entités opératoires à grandeur estimable, des construits ? Autrement dit, comme le rapportaient Fessard et Piéron, « il ne suffit pas de créer un nom, comme Intelligence mécanique, ou Aptitude musicale, pour délimiter du même coup une portion unifiée du comportement humain » (Fessard & Piéron, 1930, p. 219), comme on le ferait d'un muscle dans le système moteur d'un animal. Et si ce sont des concepts au sens strict, comment peut-on leur concevoir une grandeur et entreprendre de la mesurer ? À moins que le concept n'exclue délibérément la notion de grandeur ou de mesure et conduise alors à définir l'intelligence, par exemple, comme une « faculté spirituelle ».

Afin de faciliter la lecture, l'idée du construit, terme assez récemment traduit de l'anglais *construct*, désignera un objet mental (une construction de l'esprit) destiné à représenter quelque chose qui n'est pas explicitement concret, n'a pas en soi de grandeur et n'a de réalité que celle créée par l'opération de mesure. Par exemple, l'intelligence peut être vue comme une fiction interactionnelle et culturelle créée pour les besoins de la cause, notamment les échanges sociaux et le fonctionnement dans une culture, méritant alors la désignation de pur « concept psychométrique » et conforme avec notre conception de ce qu'est un construit.

La notion de masse corporelle, elle-même facilement mesurable grâce à l'instrumentation dont on dispose, se pose alors en parallèle à cette conception et semble plutôt pouvoir être qualifiée de trait ou encore d'attribut, en lien avec une propriété de l'objet. On s'éloigne alors de la « construction de l'esprit » pour couvrir une dimension plus matérielle et en conséquence plus objective, c'est-à-dire directement accessible dans l'objet. Toutefois, et c'est là que les choses se compliquent, il est tout à fait possible de considérer l'intelligence, mais aussi, par exemple, la motivation, la schizoïdie ou l'habileté en mathématique comme étant des propriétés de la personne plutôt que comme une fiction interactionnelle et culturelle. Sous cette optique, l'intelligence serait un trait ou un attribut, et non plus seulement un construit. Pourtant, ce choix ne la rend pas plus facile à mesurer ! Force est ainsi de constater l'émergence de multiples questions relatives au caractère des objets à mesurer en psychologie, ou en éducation, dès le début du xx^{e} siècle (voir Laurencelle & Ramsay, 2001 ; Meier, 1994). Ces questions restent d'actualité, comme il sera démontré plus loin.

Selon la supposition qu'il est possible de mesurer l'intelligence, quelle valeur accorder au score obtenu et au classement qui en découle ? Dans un article intitulé « À propos de la mesure de l'intelligence » et publié en 1904 dans *L'année psychologique*, Binet s'interrogeait sur les méthodes de mesure de l'intelligence en comparant la méthode de la cote intellectuelle à celle du degré d'instruction. Plus précisément, il cherchait à vérifier si le classement des élèves à partir d'une mesure subjective par observation (l'évaluation de l'instituteur) rendait mieux compte de l'intelligence qu'une mesure basée sur la comparaison des élèves en fonction de leur âge et du cours suivi (cours supérieur, moyen ou élémentaire) ou de la qualité de leur mémoire. Ses conclusions penchaient alors en faveur de la méthode du degré d'instruction, la considérant moins arbitraire. Binet va poursuivre ses observations destinées à convaincre les scientifiques du bien-fondé de ses constructions, mais ses différents positionnements ne permettent de trancher ni sur la question de l'existence du trait d'intelligence ni sur la question de sa mesure. Toutefois, comme l'énonçait Binet concernant la mesure de l'intelligence, « peu importent les tests, pourvu qu'ils soient nombreux » (Binet, 1910b, p. 201). Binet suggérait par cette affirmation qu'une bonne mesure de l'intelligence ne pouvait être envisagée qu'en diversifiant les mesures de celle-ci. Par cette formule, il posait peut-être les premières bases opératoires du concept de validité.

Autour de la modélisation des données

Alors que Binet tentait de convaincre ses détracteurs, Spearman (1904) se questionnait sur les aptitudes individuelles nécessaires à la résolution de problèmes plus ou moins complexes et proposait une approche factorielle, soit le modèle bifactoriel, destinée à mettre en évidence la structure du concept d'intelligence et proposant l'existence d'un facteur g comme une forme d'intelligence générale . Plus tard, Thurstone (1938) faisait évoluer la méthode en développant l'analyse multifactorielle et en contestant l'existence d'un facteur d'intelligence général. En 1952, Vernon, puis Burt en 1955 mettaient au point le premier modèle hiérarchisé à partir des modèles existants, faisant émerger des facteurs de second ordre, suivis par Horn et Cattell (1966), qui ont opérationnalisé et nommé les facteurs de second ordre « intelligence fluide », « intelligence cristallisée » et « intelligence visuospatiale ». Ainsi, certains psychométriciens voient dans cette approche la possibilité de mettre en évidence un (ou des) trait(s) à l'origine des réponses aux items du test. Les travaux de Thurstone étaient d'ailleurs réalisés dans ce sens. Toutefois, d'autres, par exemple Anastasi (1950), n'y voient qu'une modélisation mathématique permettant de regrouper des items en facteurs, sans postuler l'existence de traits sous-jacents (Angoff, 1988; Sireci, 2009). Ainsi, tout un courant de recherches repose très tôt sur la modélisation des données obtenues grâce aux tests pour légitimer les liens existants entre les traits ou construits, d'une part, et les manifestations observées, d'autre part. Pourtant, l'élaboration des instruments de collecte était également au cœur des priorités à cette époque.

Autour du développement des instruments

Au fil des travaux de la première moitié du xx^{e} siècle, les auteurs ont tenté de fournir des preuves de la validité de traits tels que la personnalité (par ex., Cattell, 1949 ; Fiske, 1949), l'intelligence (par ex., Spearman, 1904 ; Cattell, 1963) ou encore les émotions (par ex., Duffy, 1932), comme l'anxiété ou la joie. Les auteurs ont porté en priorité leurs efforts sur la définition rigoureuse des contenus des questionnaires, sur le contrôle des conditions expérimentales ou sur l'objectivité des notations, dans le but de minimiser les erreurs de mesure et de fourbir leurs instruments de mesure. La légitimité, la pertinence ou l'utilité de ces mesures étaient alors souvent reléguées au second plan, avec moins d'intérêt porté au criterium, c'est-à-dire aux corrélats observables définissant spécifi-

quement le facteur mesuré. Duffy (1932) s'est toutefois intéressée à mesurer le degré d'intensité d'une réaction émotive pour rendre compte de l'émotion. D'ailleurs, dans un article portant sur l'analyse des critères définissant l'émotion, Duffy (1934) est allée jusqu'à conclure que le concept d'émotion était sans utilité en psychologie. Elle proposait d'abandonner la catégorie « émotion » et soumettait l'idée de dimensions plus fondamentales comme le « degré d'excitation et d'inhibition ». Pourtant, la question ici n'est pas de se demander si tel auteur a raison ou a tort, mais plutôt de se questionner sur les conditions assurant l'adéquation entre ce que le test psychométrique permet de mesurer et le trait ou le construit à mesurer, peu importe que l'on parle d'émotion (un construit) ou de degré d'excitation (un trait directement mesurable). Ceci nous amène donc à suivre le fil historique des définitions successives de la validité.

L'évolution du concept de validité

Les définitions de la validité

Newton (2012) a identifié les premières traces de la notion de validité dans la documentation de l'année 1915 :

> Ainsi, Terman et al. (1915) ont discuté de « la validité d'un test d'intelligence » (p. 562) et de « la validité du QI » (p. 557) ; Starch (1916) fait référence à « la validité ou la justesse de ces mesures » (p. 3) ; Thorndike (1916) a noté « l'importance d'une échelle valide » (p. 11) ; tandis que Hartog (1918) a déploré « le fait que les tests n'ont jamais été soumis à une investigation scientifique, et que nous sommes complètement plongés dans l'obscurité en ce qui a trait à leur validité pour l'usage auquel ils sont employés » (p. 51) (Newton, 2012, p. 2, traduction libre).

Il attribue en outre la première définition des termes de validité et de fidélité à Buckingham et al. (1921), en ces termes :

> Deux des plus importants types de problèmes en mesure sont ceux qui ont trait à la détermination de ce que le test mesure, et à la consistance de cette mesure. Le premier pourrait être appelé le problème de validité et le second le problème de fidélité (Buckingham et al., 1921, p. 80, traduction libre).

Au sens originel, le problème de validité est de savoir si un test mesure réellement ce qu'il est censé mesurer (Kelley, 1927). Selon Angoff (1988), cette définition prévaut jusque dans les années 1950. Ainsi, à cette époque et pendant plusieurs décennies, la validité est définie comme le degré auquel un test mesure ce qu'il prétend mesurer. Elle exprime la qualité

externe d'un test, souvent dans une visée prédictive, par la corrélation de la mesure obtenue avec une autre mesure objective jugée pertinente (Guilford, 1946 ; Angoff, 1988). La validité étant vue initialement comme une propriété du seul test, il devient rapidement clair que la validité au sens large dépend aussi du contexte d'utilisation, du mode d'administration et peut-être même de l'usage qui est fait des résultats (Newton, 2012). Cronbach (1971) a notamment cristallisé cet élargissement de perspective en affirmant que l'on ne valide pas le test, mais l'interprétation des mesures qu'il fournit. Toutefois, la définition de la validité nécessite d'être précisée pour être opérationnalisée ; c'est l'objet des formes de validité.

L'évolution des formes de validité

Au fil du xx^{e} siècle apparaissent plusieurs formes de validité, largement utilisées et discutées, et qui ont évolué au fil du temps. Apparition et évolution correspondent d'ailleurs souvent à l'émergence ou au raffinement de techniques statistiques. Le Tableau 1 regroupe les formes les plus connues de validité, y inclus la notion de fidélité.

Même s'il est question dès 1940 de validité manifeste, qui est une forme dite naïve de validité, c'est la validité en référence à un critère, concomitante ou prédictive, qui prédomine jusque dans les années 1950. Pour Guilford et ses contemporains, ces deux formes de validité en référence à un critère correspondent à un argument sur la valeur utile du testing. Elles représentent une relation de cause à effet entre le score du test et une caractéristique mesurée à un autre moment, celui-ci pouvant être futur ou non. L'exemple qui suit offre une illustration en éducation et met en évidence une limite. L'habileté de l'élève que l'on cherche à estimer est souvent définie à partir d'un échantillonnage des performances attendues. Ainsi, dans le cas d'un test pour évaluer l'habileté à résoudre des problèmes en algèbre, le critère peut être fourni par une autre version du test, et la corrélation entre les deux versions peut offrir un argument de validité concomitante. Toutefois, la qualité de cette autre version n'est pas assurée, ce qui remet en question l'argument de validité ainsi obtenu (Kane, 2013a).

Née à la même époque, la fidélité est souvent considérée comme étant une forme de validité. Elle fait référence à la stabilité des scores relativement à diverses passations du même test (Crocker & Algina, 1986). La fidélité repose sur l'hypothèse selon laquelle il est possible de mesurer, par exemple, une activité cognitive en posant plusieurs questions dont les réponses constituent un ensemble consistant et montrent

une certaine stabilité d'un contexte à l'autre et d'un moment à l'autre (Kelley, 1942) ; le cas échéant, la fidélité repose donc aussi sur la structure interne du test (Cronbach, 1951). Ainsi, pour un questionnaire psychologique ou une batterie d'évaluation du développement moteur, un bon nombre de questions ou de tâches pour chaque aspect évalué est notamment nécessaire à l'obtention de bons coefficients de fidélité et d'une forme de validité au sens large.

À partir de 1950, la validité manifeste s'affine en validité de contenu, couvrant les facettes de l'univers à quantifier (Schmidt, 2012). Il semble opportun de noter que la validité manifeste, pourtant vue comme étant superficielle dans une vision psychométrique, est encore utilisée aujourd'hui pour valider divers instruments de collecte de données, notamment en éducation et pour la sélection de personnel (Schmidt, 2012). La validité de contenu apparaît peut-être aussi à la faveur de la création de tests dont l'objet définitionnel prêtait moins à la vérification par un critère externe, comme l'introversion/extraversion ou la déviance psychopathique. Newton et Shaw (2014) mentionnent toutefois qu'elle émerge des travaux du comité sur les tests de personnalité reposant sur une théorie qu'il remettait en question, lors de la rédaction de la version de 1954 des *Standards for educational and psychological tests* (AERA, NCME, & APA, 1954, p. 68).

À la même époque émerge la validité conceptuelle ou « de construit ». Elle peut être vue comme une validité sémantique au sens où elle consiste à repérer et à circonscrire le concept, la qualité et l'attribut reflété par la mesure en situant cette mesure dans un ensemble d'analogues sémantiques, que Cronbach et Meehl (1955) qualifient de « réseau nomologique » d'un concept psychométrique.

En 1954, la première version des *Standards for educational and psychological tests* cible trois grands types de validité qui perdureront au cours des différentes publications des standards (1966, 1974, 1985, 1999) : (1) la validité de contenu, (2) la validité conceptuelle et (3) la validité reliée à un critère.

Vers la fin des années 1970 émerge une vision de la validité reposant sur un ensemble de modèles et de méthodes (Kane, 2013a). Cette tendance se concrétise avec la validité unifiée de Messick (1989), qui inclut des considérations plus morales en lien avec les conséquences de l'usage du test et de l'interprétation des données. Ainsi, pour Messick, un bon

Tableau 1
Les formes de validité les plus connues

Années	Formes de validité	Descriptions	Auteurs clés	Approches principales
1940	Fidélité (*reliability*)		(Cronbach, 1951; Guilford, 1946; Kelley, 1942)	Alpha de Cronbach KR-20
1940	Validité manifeste (*face, apparent*)	Jugement direct (en rapport d'évidence) porté par les utilisateurs.	(Mosier, 1947; Nevo, 1985)	Pragmatique
1930-1940	Validité concomitante (en référence à un critère) (*criterion-referenced*)	Comparaison du score du test à une autre mesure de la même caractéristique (plutôt dans une approche comportementale que basée sur un trait). Les deux mesures sont prises sensiblement au même moment.	(Cureton, 1951) Voir aussi la validité pratique (Guilford, 1946)	Corrélations
1930-1940	Validité prédictive (en référence à un critère) (*criterion-referenced*)	Établissement d'un lien de prédiction entre le score du test et la mesure d'une caractéristique ou d'un comportement théoriquement associé au concept visé (idée de relation causale).	(Cureton, 1951)	Corrélations
1950	Validité de contenu (*content*)- Appelée aussi validité manifeste ou échantillonnale	Consiste à vérifier que les items correspondent à un échantillon des items possibles pour définir un domaine (idée d'univers). Souvent dans une approche déductive. Le trait est alors mis en évidence par l'homogénéité des items.	(Cureton, 1951)	Analyses factorielles. Jugements d'experts
1950	Validité de construit ou conceptuelle (*construct*) Initialement intitulée *trait validity*, elle devient *nomological validity* avec Cronbach et Meehl (Campbell, 1960)	Le test vise à mesurer un attribut ou un construit qui n'est pas défini de manière opérationnelle. Cattell (1956) proposait le terme « validation » pour parler de la validité de construit.	(Cronbach & Meehl, 1955)	Corrélations Analyses factorielles / Analyses factorielles associées à d'autres tests Matrice Multi-trait multi-méthode (Campbell & Fiske, 1959)

1990	Validité unifiée (*unified*)	La validité unifiée regroupe tous les types de validité et tient compte de l'objectif du test, de l'utilisation du score et des conséquences. Kane (2006) en propose une opérationnalisation.	(Messick, 1989)	Multiples
2000	Quelques visions actuelles	La validité unifiée doit être opérationnalisée dans une démarche de validation : elle ne l'est pas actuellement. La validité est une propriété de l'instrument qui dépend de sa sensibilité aux variations du trait mesuré. Il faut savoir comment on veut que l'instrument fonctionne et vérifier s'il fonctionne comme prévu. Réfute la validité conceptuelle. Il s'agit de relater le lien entre les observables et un attribut ou un construit théorique. Discussions sur ce qui est validité et ce qui déborde du concept.	(Kane, 2006) (Borsboom et al., 2004 ; (Scriven, 2002 ; Shadish et al., 2002) (Scriven, 2002 ; Shadish et al., 2002)	Modélisations de traits latents ou de classes latentes

Note. Les écrits proposent de multiples autres formes de validité, telles par exemple que validité convergente, factorielle, structurelle, incrémentale, discriminante, formes dont nous ne traitons pas ici en détail.

argument de validité intègre en un tout cohérent six éléments de preuve de la validité conceptuelle (ou de construit) : le contenu du test, les processus de réponse, la structure interne, les relations avec d'autres variables, la généralisation de la validité et les conséquences du testing.

Toutefois, il y a lieu de se demander, à l'instar de plusieurs auteurs (par ex., Scriven, 2002 ; Shadish, Cook, & Campbell, 2002), si cet élargissement ne déborde pas du concept de validité et ne porte pas plutôt sur la pratique professionnelle du testing et sa déontologie. En effet, pour Shadish et al., les actions, telles que la sélection de candidats, et les conséquences qui découlent du processus de mesure peuvent être évaluées, mais pas validées. Elles sont donc extérieures au processus de validation. Scriven préfère garder la définition originelle de la validité en lien avec le degré auquel un test mesure ce qu'il prétend mesurer, et reléguer ce qui touche les actions et les conséquences à ce qu'il nomme l'utilité du test. Kane (2006, 2013a, 2013b) propose ensuite un cadre de référence pour la validation, qui repose sur deux types d'arguments : des arguments de validité et des arguments liés à l'interprétation/usage.

Un siècle après les premiers écrits sur la validité, de nombreux textes et ouvrages de réflexion sont régulièrement publiés. Par exemple, le cadre de référence proposé par Kane a été largement commenté et critiqué dans les écrits, notamment dans un numéro spécial du *Journal of Educational Measurement* de 2013 (Borsboom & Markus, 2013 ; Brennan, 2013 ; Haertel, 2013 ; Moss, 2013 ; Newton, 2013 ; Sireci, 2013). De ce numéro ressortent trois catégories de commentaires. Brennan, Haertel et Moss visent à mettre en évidence les arguments liés à l'interprétation/usage et à fournir des pistes de réflexion. Newton et Sireci remettent en question l'idée de séparer les arguments en deux catégories, tous étant des arguments de validité. Finalement, Borsboom et Markus appliquent la vision de Kane à un exemple concret – celui du phlogiston – afin de démontrer que l'argumentation élaborée peut s'éloigner de la vérité et aboutir à considérer comme valide quelque chose qui ne l'est pas.

Trois livres sur la validité sont parus entre 2009 et 2014. Dans le premier, Lissitz (2009) a invité plusieurs auteurs à brosser le portrait du concept de validité selon diverses perspectives. Il propose en outre une partie axée sur des exemples pratiques d'application. Dans le deuxième, Markus et Borsboom (2013) font le lien entre les théories de la mesure et

la validité, et explorent la notion d'interprétation des scores. Finalement, dans le troisième, Newton et Shaw (2014) proposent un historique très complet du concept de validité et des éléments de réponses à une diversité de questions en lien avec la validité ou la validation.

Ainsi, de nouvelles visions continuent à émerger. Leur objectif est, dans certains cas, de rapporter la validité à des propriétés plus psychométriques ; dans d'autres, de revenir au concept de validité tel qu'il a initialement été défini et de le rendre moins généraliste ; et, enfin, d'élargir le concept. Les définitions, les formes de validité et les manières de les opérationnaliser continuent donc à faire réfléchir autant les praticiens que les chercheurs et à susciter des débats. Dans ce qui suit, un intérêt particulier sera porté au lien entre validité et théories de la mesure, en passant par une articulation à la notion de causalité.

La validité, les théories de la mesure et la causalité

Logiquement, avant de poser la question de la validité d'un test ou d'une mesure, il est essentiel d'assurer que l'on mesure *quelque chose*, que ce quelque chose soit un trait objectivement réel ou un pur construit. L'exemple de l'intelligence sera repris ici pour cadrer les propos et pour mettre en évidence diverses manières d'aborder la validité, ainsi que les postulats sur lesquels elles reposent.

Ainsi, selon une posture qui place le trait d'intelligence dans un réseau de relations causales, l'intelligence est censée exister comme un trait possédé par la personne et expliquer la réponse fournie aux items des tests de QI. En revanche, dans une posture acausale, ce sont les items des tests de QI, soit un échantillon d'items possibles définissant un domaine, qui génèrent un ensemble de réponses dont on désigne la valeur par «l'intelligence», le degré d'intelligence. Dans le premier cas, la causalité est centrale et émane d'un trait, alors que, dans le second cas, c'est une généralisation qui distille le construit. Ces différentes postures impliquent des analyses de natures différentes et suggèrent des conceptions différentes de la validité (Markus & Borsboom, 2013).

Pour tenter de clarifier ce point, reprenons un exemple proposé par Borsboom (2006), où l'on souhaite construire et valider un instrument de mesure d'un trait de personnalité (*être consciencieux*). Trois manières de voir les choses sont possibles :

1- Les items de l'instrument sont un échantillon d'items-situations possibles définissant le domaine correspondant à *être consciencieux.* La proportion des items maîtrisés ou endossés offre alors une possible mesure de ce trait de personnalité. Dans ce cas, c'est le domaine, conceptuellement rassemblé, qui définit le trait.

2- Les comportements ciblés et sous-tendus par les items-situations causent ce qu'on appelle *être consciencieux.* Les items renvoient à une collection de comportements morcelés, lesquels sont alors conceptualisés en un tout pour former le trait.

3- Le fait d'*être consciencieux* cause la manière de répondre aux items-situations. La personne possède le trait et ses réactions aux items en reflètent la force.

Dans les deux premières manières de voir les choses, la théorie de référence sera la théorie classique des tests, laquelle suppose l'existence d'un score vrai relatif au trait duquel on souhaite s'approcher.

Dans le premier cas, pour valider l'instrument, la théorie classique des tests propose l'étude de la cohérence interne du contenu du test, laquelle repose souvent sur le calcul du coefficient alpha de Cronbach, qui reflète simplement l'intercorrélation des items selon les dimensions attendues. Ce coefficient, qu'on associe aussi à la fidélité du test, est particulièrement important dans les études de généralisabilité (Laveault, 2012) qui sont préconisées dans l'étude de la validité, selon cette manière de voir les choses. Ces études de généralisabilité permettent de tenir compte des nombreuses sources d'erreur et de biais possibles dans l'estimation de la fidélité (Laveault & Grégoire, 2014). En effet, la théorie de la généralisabilité permet d'étudier dans quelle mesure les différentes facettes (items, élèves, correcteurs, par exemple pour un test scolaire) affectent les mesures. Ainsi, l'établissement d'un degré de consistance permet de juger si le test repose sur un échantillonnage satisfaisant du concept visé, et si la mesure résultante est valide en ce sens restreint. Également, il existe d'autres manières d'estimer la fidélité (Lord & Novick, 1968), sans référence cette fois avec la cohérence du test ou son interprétation.

Dans le deuxième cas, les comportements ciblés permettent de conceptualiser le trait. Les analyses pointent alors vers des modèles formatifs visant à mettre en évidence des variables composites (voir la Figure 1) avec, par exemple, des analyses factorielles d'un mode ou d'un autre. La mise en évidence du trait (ou du construit) dépend alors de manière causale des réponses aux items. Par exemple, l'ensemble des bonnes et mauvaises réponses d'un élève à un test en algèbre permet d'en inférer son niveau d'habileté.

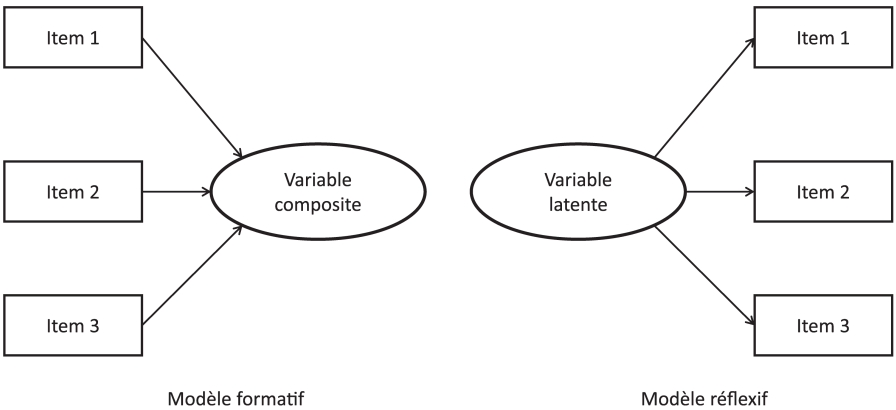

Figure 1. ***Illustration des deux types de relations causales***

La préoccupation de fidélité est compatible avec l'idée de mettre en évidence un trait ou un construit mesuré par le test, mais sans nécessairement permettre de l'identifier. Certains auteurs utilisent des modèles d'équations structurelles pour estimer la propriété de fidélité. Par exemple, Rindskopf et Rose (1988) mettent en évidence la portion de la variance vraie qui est en lien avec la mesure visée, permettant ainsi de calculer un indice de fidélité et de vérifier l'existence d'un contenu mesurable.

L'étude de la dimensionnalité du test, ou de la manière dont chaque item est en lien avec chaque dimension du test, est une procédure commune dans les études de validation inscrites dans ce deuxième cas. Elle repose fréquemment sur des analyses factorielles exploratoires ou confirmatoires, au gré des assises théoriques ou des préférences techniques dont s'inspirent les auteurs. Ces modélisations s'inscrivent plus largement dans la famille des modèles d'équations structurelles (voir par ex.

Bollen, 1989 ; Kline, 2011). Ces diverses procédures, impliquant variables manifestes et latentes, peuvent viser la vérification de l'unidimensionnalité de l'instrument ou de chacune de ses sections, mais également la validité de construit dans une approche formative ; la validité convergente par la quantification des liens entre les items et le concept (trait ou construit) ; la validité discriminante par la vérification de la spécificité des concepts représentés par les variables latentes ; et, plus difficilement, la validité nomologique de l'outil au sens de la concordance entre ce que mesure l'outil et la réalité étudiée (Whitely, 1983).

Dans le troisième et dernier cas, l'hypothèse de l'existence d'une variable latente (*être consciencieux*), représentant un trait ou un construit causant les réponses aux items, pointe vers des modèles réflexifs (voir la Figure 1). Dans ce cas, c'est par exemple l'habileté d'un élève en algèbre qui cause sa manière de répondre à un item, et le modèle met en relation l'habileté avec la probabilité de répondre convenablement à chaque item. La théorie de réponse à l'item, développée à partir des années 1950, offre des modèles de mesure compatibles avec la vision réflexive des liens entre la variable latente (construit ou trait de nature continue ou discrète) que l'on cherche à mesurer et les variables observées.

À cet égard, Borsboom, Mellenbergh et van Heerden (2004) en arrivent à une définition de la validité qui s'éloigne des préoccupations liées à l'épistémologie, à la définition des concepts, à la corrélation et, de manière générale, aux formes de validité présentées dans le Tableau 1. Pour ces auteurs, toutes les démarches apparentées aux formes de validité sont des procédures pour la validation, mais les concepts de validité et de validation ne peuvent pas être utilisés de manière interchangeable (Borsboom et al., 2004). Ils réfutent ainsi l'idée, pourtant largement acceptée depuis au moins deux décennies, que la validité a à voir avec l'interprétation des scores. Leur conception de la validité repose simplement sur la manière dont l'instrument est capable de capter les variations de l'attribut qui est l'objet de la mesure. Pour définir la validité, ils s'appuient sur deux postulats : (1) l'existence théorique de l'attribut ou du construit qu'ils veulent mesurer, et (2) la relation théorique de cause à effet entre les variations de l'attribut (ou du construit) et les réponses aux items. Dans cette approche, la validité concerne essentiellement l'élaboration d'une théorie et d'une preuve reliant le processus de réponse et l'objet à mesurer. L'utilisation de modèles de traits latents, tels que ceux issus de la théorie de réponse à l'item (TRI), incluant éventuellement des

études de fonctionnement différentiel d'item (DIF) (Markus & Borsboom, 2013), offre alors un moyen de mettre en perspective les données empiriques et la théorie en vérifiant notamment les propriétés de mesure des instruments ou la prévalence de certains patrons de réponses théoriquement attendus. Toutefois, force est de constater que le discours de Markus et Borsboom reste souvent théorique, c'est-à-dire peu pragmatique.

À l'heure actuelle, il est courant d'utiliser un modèle de Rasch (1960) pour valider les propriétés de mesure d'un instrument (par ex., Pallant & Tennant, 2007). Il est alors question d'étudier sa fidélité et sa capacité à discriminer plutôt que sa validité. Pourtant, ces analyses sont souvent réalisées en complément à des analyses factorielles qui, elles, visent l'étude de la validité conceptuelle. Cette combinaison des approches tend à cumuler des éléments de preuves de validité de diverses natures, combinant plusieurs des visions présentées ci-dessus, et elle exprime ainsi une démarche de validation.

Pour résumer, le concept à cerner pourrait être la condition préalable à l'expérience ou le résultat de l'expérience. Pour revenir à l'exemple de l'intelligence, son étude et sa mesure peuvent se baser sur l'hypothèse selon laquelle l'intelligence existe et qu'on en a une certaine idée. Vérifier cette hypothèse s'inscrit dans la perspective d'un modèle réflexif. Toutefois, l'intelligence peut tout autant n'être simplement que le concept d'un objet fictif, une construction mentale existant seulement dans notre intellect. Il s'agit alors de mesurer d'abord et de définir ensuite, et c'est ce que propose le modèle formatif. Le risque dans le premier cas est peut-être d'introduire trop de subjectivité dans les connaissances scientifiques et, dans le second, de priver les concepts de tout élément subjectif et de ne laisser aucune place à l'expérience.

Ainsi, selon l'approche retenue, la validité fait référence soit à la définition du domaine et à la représentativité des items qui le constituent dans une approche plutôt centrée sur le contenu, soit à la définition du trait et à l'étude des liens de causalité sans accent particulier sur le contenu (Markus & Borsboom, 2013). Il semble difficile de statuer si l'une ou l'autre des postures est plus appropriée, et même si elles sont éventuellement conciliables. Ce tour d'horizon des conceptions de la validité met en évidence la variété des acceptions et applications possibles du concept de validité, variété qui sera illustrée par quelques exemples issus du domaine de l'éducation.

Des réflexions à partir de quelques exemples d'application en éducation

Cet exposé sur les formes de validité et la réflexion qu'il engendre sur les liens entre validité, causalité et mesure ont permis de mettre en évidence la diversité et la complexité des postures définitionnelles existantes. Cette diversité et cette complexité expliquent probablement pourquoi le choix des auteurs de nombreuses études de validité reste souvent obscur ou implicite ; semble parfois arbitraire, voire expéditif, et laisse quelquefois au lecteur le sentiment que d'autres choix auraient pu être faits.

Lorsqu'il s'agit de valider l'utilisation d'un instrument, assez fréquentes sont les procédures de validation « clé en main », c'est-à-dire des recettes à appliquer. En éducation, la validité unifiée proposée par Messick (1989) est un modèle souvent privilégié dans le cadre de la validation des questionnaires d'intérêts, impliquant une extension de la validité aux conditions d'utilisation de l'instrument et à ses conséquences pour la personne testée ou son organisation. Par exemple, Hébert (2013) opérationnalise les six éléments de preuve de validité de Messick pour vérifier si une épreuve ministérielle en mathématique visant à évaluer les compétences des élèves à la fin du troisième cycle de primaire est valide. Elle comptabilise ensuite les éléments de preuve selon qu'ils pointent vers le caractère valide de l'épreuve ou pas. Finalement, les conclusions de l'étude d'Hébert remettent en question autant la pertinence des six éléments de preuve proposés par Messick que la validité de l'épreuve en question. C'est une démarche similaire que proposent Voss, Kunter et Baumert (2011) lorsqu'ils cherchent à valider un test de connaissances générales en pédagogie et en psychologie chez les enseignants en s'inspirant de la version de 2004 des *Standards for educational and psychological testing* et des propositions de Messick (1989). Ces auteurs ne parviennent pas à conclure sur les qualités psychométriques du test et suggèrent de fournir d'autres preuves de validité.

Dans le domaine de l'éducation, les objectifs poursuivis concernent souvent soit la mesure des différences interindividuelles, soit la mesure des apprentissages. Ces deux objectifs seront illustrés en évoquant la validation de deux questionnaires en éducation, l'un destiné à évaluer les perceptions sur l'évaluation formative des apprentissages et l'autre dont l'objectif est de repérer les élèves à risque de décrochage.

Dans le premier cas, Pat-El, Tillema, Segers et Vedder (2013) s'étaient donné pour objectif de valider deux questionnaires sur l'évaluation formative, l'un destiné aux enseignants et l'autre aux élèves. Sur la base d'analyses factorielles confirmatoires (AFC), ces auteurs ont mis en évidence la nature robuste de la structure factorielle des deux questionnaires, et ont conclu que ceux-ci étaient valides et permettaient de juger la contribution de l'évaluation à l'apprentissage. Même si l'étude est annoncée comme étant une étude de validation, l'AFC n'a servi qu'à confirmer des structures sémantiques organisant les éléments d'un modèle, sans apporter aucune preuve de la validité de l'outil. À aucun moment, les auteurs ne font explicitement le lien entre leurs résultats et le concept de validité, et leurs analyses ne renseignent pas sur les progrès des élèves. Ces questionnaires n'ont pas été construits de manière à être sensibles aux progrès ou, en tout cas, ils n'ont pas été validés pour cet usage. Cet exemple est illustratif d'un certain nombre d'études de validité qui n'incluent aucune réflexion sur ce qu'est la validité ni sur le lien qu'il y a entre les analyses des données empiriques proposées et des preuves de validité, selon l'une ou l'autre conception ou forme de validité.

Comparativement à cet exemple, la validation d'un questionnaire de dépistage (Potvin et al., 2010) satisfait à l'objectif des auteurs, qui était de permettre un repérage des élèves à risque de décrochage selon quatre types de caractéristiques : « comportements antisociaux cachés », « peu intéressé/peu motivé », « problème de comportement » et « dépressif ». Afin de satisfaire à l'objectif, ces auteurs ont testé plusieurs formes de validité, soit la validité en référence à un critère afin de montrer le lien entre l'instrument et le critère « décrocheur », et la validité conceptuelle (ici, de type diagnostique) en montrant que les élèves à risque se distinguent de façon appropriée de ceux qui ne le sont pas.

Quels que soient l'objectif de l'entreprise de validation et la forme technique de la procédure de validation mise en œuvre, les preuves de validité (s'il est permis d'utiliser ce terme de preuve) devraient rendre explicite le lien entre ce qui est concrètement démontré par la procédure et l'objectif de la validation. La corrélation, l'analyse factorielle, la modélisation en équations structurelles, la TRI et la régression multiple ne sont que des techniques mathématiques et ne peuvent à elles seules constituer des arguments ni des preuves. La preuve elle-même doit être explicative plutôt qu'assertorique ; rendre explicite et justifier l'articulation entre les résultats de la procédure appliquée et la conclusion qui s'ensuit ; et, enfin, puisqu'il

s'agit d'un travail scientifique et non partisan, faire état des conditions limitatives qui s'appliquent. Une fois cette preuve convenablement présentée, la question de savoir de quelle sorte ou forme de validité il s'agit devient éventuellement secondaire.

Conclusion et perspectives

Finalement, force est de constater que la validité d'un test psychométrique peut être questionnée pour plusieurs raisons : prouver la réalité même d'un concept et l'existence de son substrat ; clarifier sa nature, sa teneur sémantique ; classer des individus justement ; prédire des résultats ou des conséquences ; ou quantifier la valeur d'une performance. Au fil du temps, de multiples définitions et visions du concept de validité ont nourri quantité d'écrits. Malgré la variété des points de vue et des méthodes, il se dégage des textes parcourus l'idée générale que la validité doit être syntonisée aux objectifs poursuivis par les utilisateurs et que les méthodes pour ce faire doivent être choisies en conséquence. Quant au concept de validité unifiée de Messick (1989), tout séduisant soit-il, il n'a pas donné les fruits escomptés (Markus & Borsboom, 2013 ; Scriven, 2002 ; Shadish et al., 2002) et reste à être lui-même validé. D'ailleurs, Sussmann et Robertson (1986), après analyse de plusieurs designs permettant de mettre en œuvre une démarche de validation, concluent à la nécessité de diversifier les designs selon les objectifs visés, ce qui semble incompatible avec une vision unifiée de la validité.

Établir la validité d'un test ou d'un instrument de mesure signifie en pratique que l'on a mis à contribution une méthode, un processus de validation. À rebours, le processus de validation employé va contribuer à camper et à définir de façon opératoire la validité du test. Validité et processus de validation devraient donc, en principe, être accordés l'un à l'autre, mais le sont-ils toujours ? Et lesdites méthodes de validation (corrélations, analyses factorielles exploratoires ou confirmatoires, modèles structuraux, analyses discriminantes, alpha de Cronbach, régressions et analyses acheminatoires, etc.) semblent ne pas être toujours judicieusement appliquées et interprétées dans les études publiées. Il reste aussi la question de savoir si toutes ces analyses font partie de la mise en évidence de la validité, ou si elles en sont exclues. Elles peuvent permettre, en accord avec la vision proposée par Borsboom (2006), de peaufiner une théorie à mettre empiriquement à l'épreuve par la suite.

RÉFÉRENCES

American Educational Research Association, National Council on Measurement in Education, & American Psychological Association (1954). *Technical recommendations for psychological tests and diagnostic techniques.* Washington, DC: Washington.

American Educational Research Association, National Council on Measurement in Education, & American Psychological Association (1966). *Standards for educational and psychological tests and manuals*. Washington, DC: American Psychological Association.

American Educational Research Association, National Council on Measurement in Education, & American Psychological Association (1974). *Standards for educational and psychological tests*. Washington, DC: Washington.

American Educational Research Association, National Council on Measurement in Education, & American Psychological Association (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

American Educational Research Association, National Council on Measurement in Education, & American Psychological Association (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

American Educational Research Association, National Council on Measurement in Education, & American Psychological Association (2004). *Standards for educational and psychological testing* (2nd ed.). Washington, DC: American Educational Research Association.

Anastasi, A. (1950). The concept of validity in the interpretation of test scores. *Educational and Psychological Measurement, 10*(1), 67-78. doi:10.1177/001316445001000105

Angoff, W. H. (1988). Validity: An evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test validity* (p. 19-32). Hillsdale, NJ: Routledge.

Binet, A. (1904). À propos de l'intelligence. *L'année psychologique, 11*(11), 69-82. doi:10.3406/psy.1904.3667

Binet, A. (1910a). Qu'est-ce qu'une émotion? Qu'est-ce qu'un acte intellectuel? *L'année psychologique, 17*(1), 1-47. doi:10.3406/psy.1910.7270

Binet, A. (1910b). Nouvelles recherches sur la mesure du niveau intellectuel chez les enfants d'école. *L'année psychologique, 17*(17), 145-201.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71*(3), 425-440. doi:10.1007/s11336-006-1447-6

Borsboom, D., & Markus, K. A. (2013). Truth and evidence in validity theory. *Journal of Educational Measurement, 50*(1), 110-114. doi:10.1111/jedm.12006

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071. doi:10.1037/0033-295X.111.4.1061

Brennan, R. L. (2013). Commentary on "Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 74-83. doi:10.1111/jedm.12001

Buckingham, B. R., McCall, W. A., Otis, A. S., Rugg, H. O., Trabue, M. R., & Courtis, S. A. (1921). Report of the standardization committee. *Journal of Educational Research, 4*(1), 78-80.

Burt, C. (1955). The evidence for the concept of intelligence. *British Journal of Educational Psychology, 25*(3), 158-177. doi:10.1111/j.2044-8279.1955.tb03305.x

Cattell, R. B. (1949). The dimension of culture patterns by factorization of national characters. *Journal of Abnormal and Social Psychology, 44*(4), 443-469. doi:0.1037/h0054760

Cattell, R. B. (1963). Theory of fluid and cristallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*(1), 1-22. doi:10.1037/h0046743

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovitch.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-335. doi:10.1007/BF02310555

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302. doi:10.1037/h0040957

Duffy, E. (1932). The measurement of muscular tension as a technique for the study of emotional tendencies. *American Journal of Psychology, 44*(1), 146-162. doi:10.2307/1414961

Duffy, E. (1934). Emotion: An example of the need for reorientation in psychology. *Psychological Review*, *41*(2), 184-198. doi:10.1037/h0074603

Fessard, A., & Piéron, H. (1930). La notion de validité. *L'année psychologique, 31*(1), 217-228. doi:10.3406/psy.1930.30008

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology, 44*(3), 329-344. doi:10.1037/h0057198

Guilford, J. P. (1946). New standards for test evaluation. *Educational and Psychological Measurement, 6*, 427-439. doi:10.1177/001316444600600401

Haertel, E. (2013). Getting the help we need. *Journal of Educational Measurement, 50*(1), 84-90. doi:10.1111/jedm.12002

Hébert, M.-H. (2013). *Validation d'une épreuve pour rendre compte du niveau de développement des compétences du programme de mathématique pour l'enseignement primaire* (thèse de doctorat non publiée). Québec, Canada : Université Laval.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*(5), 253-270. doi:10.1037/h0023816

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement, (*4th ed.; p. 17-64). Westport, CT: American Council on Education and Praeger.

Kane, M. T. (2013a). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1-73. doi:10.1111/jedm.12000

Kane, M. T. (2013b). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement, 50*(1), 115-122. doi:10.1111/jedm.12007

Kelley, T. L. (1927). *Interpretation of educational measurements*. New York, NY: Macmillan.

Kelley, T. L. (1942). The reliability coefficient. *Psychometrika, 7*(2), 75-83. doi: 10.1007/BF02288068

Kline, R. B. (2011). *Principles and practice of structural equation modeling* (3rd ed.). New York, NY: Guilford Press.

Laurencelle, L., & Ramsay, J. O. (2001). À la recherche de l'« unité de mesure » en psychométrie : réflexions sur la mesure en sciences humaines. *Mesure et évaluation en éducation, 24*, 41-52.

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

Laveault, D., & Grégoire, J. (2014). *Introduction aux théories des tests en psychologie et en sciences de l'éducation* (3e éd.). Louvain-la-Neuve, Belgique: De Boeck.

Lissitz, R. W. (2009). *The concept of validity*. Charlotte, NC: IAP.

Lord, F. M., & Novick, M. R. (1968). *Statistical theory of mental test scores*. Massachusetts, MA: Addison Wesley.

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

Meier, S. T. (1994). History. In S. T. Meier (Ed.), *The chronic crisis in psychological measurement and assessment* (p. 1-33). San Diego, CA: Academic Press.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (p. 13-103). Washington, DC: American Council on Education and Macmillan.

Moss, P. A. (2013). Validity in action: Lessons from studies of data use. *Journal of Educational Measurement, 50*(1), 91-98. doi:10.1111/jedm.12003

Newton, P. E. (2012). Clarifying the consensus definition of validity. *Measurement: Interdisciplinary Research and Perspectives, 10*(1-2), 1-29. doi:10.1080/15366367.2012.669666

Newton, P. E. (2013). Two kinds of argument? *Journal of Educational Measurement, 50*(1), 105-109. doi:10.1111/jedm.12004

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational and psychological assessment*. Cambridge, UK: Sage Publications.

Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An exemple using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46*(1), 1-18. doi:10.1348/014466506X96931

Pat-El, R. J., Tillema, H., Segers, M., & Vedder, P. (2013). Validation of assessment for learning questionnaires for teachers and students. *British Journal of Educational Psychology*, *83*(1), 98-113. doi:10.1111/j.2044-8279.2011.02057.x

Potvin, P., Doré-Côté, A., Fortin, L., Royer, E., Marcotte, D., & Leclerc, D. (2010). *Questionnaire de dépistage d'élèves à risque de décrochage scolaire.* Québec, Canada : Centre de transfert pour la réussite scolaire au Québec. http://www.pierrepotvin.com/8.%20Banque%20d%27outils/questionnaire-de-depistage.pdf

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research, 23*(1), 51-67. doi:10.1207/s15327906mbr2301_3

Schmidt, F. L. (2012). Cognitive tests used in selection can have content validity as well as criterion validity: A broader research review and implications for practice. *International Journal of Selection and Assessment*, *20*(1), 1-13. doi:10.1111/j.1468-2389.2012.00573.x

Scriven, M. (2002). Assessing six assumptions in assessment. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (p. 268-287). Mahwah, NJ: Erlbaum.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inferences*. Boston, MA: Houghton Mifflin.

Sireci, S. G. (2009). Packing and unpacking sources of validity evidence: History repeats itself again. In R. W. Lissitz (Ed.), *The concept of validity* (p. 19-37). Charlotte, NC: IAP.

Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement, 50*(1), 99-104. doi:10.1111/jedm.12005

Spearman, C. (1904). "General intelligence, objectively determined and measured. *American Journal of Psychology, 15*(2), 201-292. doi:10.2307/1412107

Sussmann, M., & Robertson, D. U. (1986). The validity of validity: An analysis of validation study designs. *Journal of Applied Psychology, 71*(3), 461-468. doi:10.1037/0021-9010.71.3.461

Thurstone, L. L. (1938). *Primary mental abilities*. Chicago, IL: University of Chicago Press.

Vernon, P. E. (1952). *La structure des aptitudes humaines*. Paris, France : PUF.

Voss, T., Kunter, M., & Baumert, J. (2011). Assessing teacher candidates' general pedagogical/psychological knowledge: Test construction and validation. *Journal of Educational Psychology, 103*(4), 952-969. doi:10.1037/a0025125

Whitely, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93*(1), 179-197. doi:10.1037/0033-2909.93.1.179