

La validité psychométrique est d'abord un jugement : une perspective pragmatique

Nathalie André ^a   and Louis Laurencelle ^b 

^aCentre de Recherches sur la Cognition et l'Apprentissage, Université de Poitiers

^bUniversité du Québec à Trois-Rivières

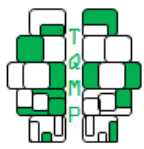
Abstract ■ Validité, validation d'un test, propriétés ou qualités psychométriques, voilà trois expressions qui désignent les opérations ou procédures permettant de juger de la pertinence d'un instrument de mesure psychométrique, de son élaboration et son application. L'expression "validité psychométrique" a dès les débuts reçu diverses interprétations et elle continue aujourd'hui d'interpeller les théoriciens. La validité d'un test réside-t-elle dans la mesure qu'il produit ou dans l'action qui en découle ou bien réfère-t-elle à l'interprétation qu'on lui attache? La validité repose-t-elle strictement sur le résultat d'une procédure de validation? Un test "démonstré valide" est-il valide en lui-même ou sa "validité" est-elle relative à la démonstration qu'on en a faite ou à son usage prévu? Existe-t-il différentes espèces de validité ou représente-t-elle un concept unifié, molaire? Une définition claire et consensuelle du concept de "validité psychométrique" n'a pas encore vu le jour. C'est dans une perspective à la fois pragmatique et conceptuellement rigoureuse que nous préconisons d'abolir l'expression "validité psychométrique", que cette expression ne correspond pas *per se* à un concept, mais qu'elle est tout simplement un jugement, tout comme l'est au tribunal le prononcé du juge en considération des preuves qu'on lui soumet, de l'intimé concerné et des circonstances atténuantes au procès. La validité est un jugement à l'effet que la mesure, le test, l'échelle psychométrique répond adéquatement et fidèlement à l'usage qu'on lui destine. L'essai explore ensuite le riche arsenal disponible au psychométricien et à l'utilisateur pour leur permettre "d'évaluer" à bon escient la validité situationnelle du test produit. ■ Validity, test validation, psychometric properties or qualities: here are three expressions that designate the operations and procedures that make it possible to judge the relevance of a psychometric measurement instrument, its development and its use. The expression "psychometric validity" has historically inherited various interpretations from the outset and it continues today to challenge theorists. Does the validity of a test lie in the measure it produces or in the action that is derived from it, or does it refer to the interpretation intended for it? Is validity based strictly on the result of a validation procedure? Is a "demonstrated valid" test valid in itself or is its "validity" relative to the demonstration or intended use? Are there different species of validity or does it represent a unified, molar concept? A clear and consensual definition of the concept of "psychometric validity" has not yet been found. It is from a pragmatic yet conceptually rigorous perspective that this essay argues for the abolition of the term "psychometric validity", that the term is not a concept *per se*, but simply a judgment, just as a judge's pronouncement in court is a judgment in light of the evidence before him or her, the respondent, and the circumstances surrounding the trial. Validity is a judgment that the measure, test, or psychometric scale responds adequately and reliably to its intended use. The essay then explores the well-stocked arsenal available to the psychometrician and his client to enable them to "assess" themselves the situational validity of the test produced.

Keywords ■ Validité, Validation psychométrique, Modèles factoriels, Modèles à facettes de Guttman
■ Validity, Psychometric test validation, Factorial models, Guttman's facet models.

 nathalie.andre@univ-poitiers.fr

 [10.20982/tqmp.18.3.p224](https://doi.org/10.20982/tqmp.18.3.p224)

Acting Editor ■ Denis Cousineau (Université d'Ottawa)



Introduction

Cet essai se situe dans un débat qui remonte à l'origine de la mesure en psychologie et des tests psychométriques, débat qui a connu quelques marées montantes dans le passé et qui gronde aujourd'hui comme jamais, celui de la validité des tests psychométriques (ANDRÉ, LOYE & LAURENCELLE, 2015). En métrologie générale, "valider un instrument de mesure" revient essentiellement à l'étalonner (BIPM, 2012; CHEVALIER, CONTE, LABRIE, LANDRY & NADEAU, 1978), le procédé consistant à construire un diagramme ou courbe d'étalonnage mettant en relation la grandeur (physique) évaluée et la mesure qui en est prise. Les propositions faites sur la validité des instruments psychométriques remontent notamment à KELLEY (1927) et GUILFORD (1946) et réfèrent tour à tour à trois considérations : la signification (ou interprétation) apparente des mesures, la relation de la mesure à un critère quantifiable, la fiabilité ou fidélité de la mesure produite. Comme le montrent ANDRÉ et collègues (2015) dans leur bilan de la littérature sur la validité psychométrique, le concept a ensuite évolué en bifurquant en deux branches, l'une concernant la définition formelle de la validité (BORSBOOM, MELLEBERGH & van HERDEN, 2004), l'autre axée sur les procédures de validation (MESSICK, 1989; KANE, 2006, 2013a, 2013b). Or, par contraste avec l'étalonnage d'un instrument de mesure physique, la validation d'un instrument psychométrique n'a généralement pas de grandeur physique de référence, de critère direct, ce qui a amené les auteurs à proposer le terme "construit psychométrique", en anglais *psychometric construct*, donné comme substitut à cette grandeur (CRONBACH & MEEHL, 1955). Les approches plus récentes concernent principalement la mesure des caractéristiques 'abstraites' de la personne. On y retrouve des systèmes d'évaluation théoriques, tels ceux de Kane (2006, 2013a) et de Mislevy et ses collaborateurs (MISLEVY, ALMOND & LUKAS, 2003; ZIEKY, 2014), systèmes qui insistent à juste titre sur la preuve et en déclinent l'application à travers une série de volets touchés par l'application d'un test. LOYE (2018) présente une étude fine et bien documentée de ces approches, leurs différences et leur possible complémentarité.

Qu'est-ce que la validité, peut-on démontrer qu'un test est valide, comment doit-on procéder pour la validation d'un test ? Voilà trois ordres de questions que nous aborderons à tour de rôle dans cet essai, ce dans un souci de rigueur terminologique comme de pragmatisme pour le développeur de tests, notre espoir étant d'évacuer un peu du brouillard, voire de la confusion qui couvre la praxéologie de la validation psychométrique.

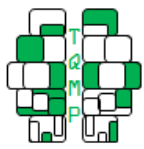
Qu'est-ce que la validité psychométrique ?

Peut-on définir la validité psychométrique ? "Valider", dans le dictionnaire, équivaut à "déclarer propre à sa fonction", et quelques analogues proposés en sont "ratifier", "sanctionner", "homologuer". "Valider" réfère donc expressément à un jugement porté sur la chose, l'action, le résultat en question, jugement donné par une instance qui en a l'autorité ou en vertu d'une règle convenue appliquée sur la satisfaction de conditions et de critères définis. Dans le cas d'une mesure physique comme le pèse-personne, valider se ramène à établir la fonction d'étalonnage devant rencontrer des critères de justesse et de précision qui lui sont propres. Adressée à la mesure d'une "grandeur psychologique", un construit, la sanction de validité reste un jugement, un jugement humain, un verdict basé sur un ensemble contextuel de règles, soit le résultat d'une procédure de validation, et assumé par une personne ou par un groupe de personnes, selon le contexte spécifique de la mesure. Dans cette perspective, la validité n'équivaut pas à la somme des résultats de validation : elle est un jugement, opérationnalisé par des règles et argumenté, partiellement arbitraire et intrinsèquement contestable. De là, il ressort que l'effort de construire une définition substantielle et consensuelle de la validité psychométrique est, selon nous, vain et voué à l'échec. Une définition de la validité ne s'appliquera jamais à toutes les combinaisons d'instruments et d'usages à valider.

En résumé, la fonction du 'valideur' d'un test ou d'une mesure consiste à choisir et appliquer la procédure de validation (comportant une ou plusieurs applications d'une ou de plusieurs méthodes) qu'il croit appropriée, en argumenter le résultat tout en explicitant minutieusement ce que cette procédure démontre, et expliquer en quoi (sous quel aspect) et comment (par quel raisonnement et selon quel jeu de résultats) cela valide la mesure ou le test pour l'usage qui en est prévu. Quant à l'utilisateur ou à l'instance de sanction normative, après avoir considéré le test recommandé et son objet (but ou dessein déclaré, application envisagée), il lui faut juger de la valeur de la preuve fournie, décidant alors si l'usage de la mesure qu'il en prévoit est ou n'est pas valide.

Que signifie "Le test ABC est valide" ?

Pour l'usager d'un test psychométrique, pour la personne mesurée, que signifie la phrase : "Le test ABC est valide" ? En marge des discours savants et du jargon de la métrologie et de la psychométrie, on comprendra qu'un test, un instrument de mesure, est valide s'il renseigne adéquatement sur une caractéristique, un état, une "grandeur" à laquelle il est censé convenir. En ce sens



très général, “valide” est synonyme de “bon”, d’“utile”, d’“approprié”, de “vérifiable”. Poussant un peu plus loin l’examen lexicologique du lexème “valide” tel qu’appliqué à un instrument de mesure psychométrique, on dira que, pour mériter cette épithète dans un contexte donné, la mesure, comme l’instrument qui la porte, doit être :

fidèle : à valeur répétable pour la même grandeur d’un moment à l’autre, d’un évaluateur à l’autre;

représentative : présentant en transposition (par réduction ou symbolisation) la valeur caractéristique de la grandeur visée;

sensible : réactive aux variations de grandeur ou d’état dans l’objet ou la personne mesurée;

discriminante : ayant une unité de mesure assez fine pour bien séparer les grandeurs, ou un pouvoir de catégorisation de ces grandeurs suffisant (degrés d’anxiété, niveaux d’intelligence,...);

univoque : ne reflétant que la seule caractéristique ou grandeur visée de la personne;

structurellement définie : dont la base du score et sa méthode de calcul sont explicites, justifiées et logiques;

d’interprétation simple : dont l’interprétation est directe (ou évidente) ou indirecte via les conséquences certaines ou probables liées à la mesure;

conceptuellement ou objectivement identifiée : dont soit l’identification conceptuelle, soit la référence (externe) objective, soit les deux sont établies et claires;

adéquatement normée : dans le cas des applications cliniques (sélection, diagnostic, qualification), par rapport à une population ou sous-population spécifiée et prenant en compte la dérive ou l’obsolescence possible des niveaux normatifs;

à conclusion prudente et décision imputable : dans le cas notamment des applications cliniques et de qualification, avec prise en compte explicite des erreurs de mesure et de la spécificité des contextes;

utilitaire (axée sur et définie par son utilité objective) : dont l’utilité pratique pour les applications prévues de la mesure a été démontrée (dans le cas des grilles prédictives : voir plus loin).

La validité dans les publications scientifiques

Si on consulte maintenant les publications d’ordre scientifique sur le sujet, c.-à-d. les manuels ou traités de psychométrie, les textes savants ou théoriques, les articles surtout qui “valident” un test psychométrique, on constate qu’il existe un certain flou sur la définition du concept de validité, flou que nous croyons pouvoir élucider en trois aspects. Ainsi, selon les usages, un test sera dit valide si ...

1. on a pu montrer que les données récoltées de l’instrument covarient avec des grandeurs reflétant directement ou associées théoriquement à la grandeur visée;
2. on a pu montrer que les données récoltées de l’instrument

ont une structure (statistique) compatible avec un modèle théorique préétabli;

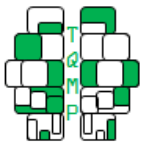
3. une procédure ou quelques procédures dite/s de validation appliquée/s à l’instrument est/sont “concluante/s”.

Les deux premières catégories sont, on en conviendra, liées à l’argument de validité en ce qu’elles recoupent plusieurs des connotations du mot “valide” énumérées plus haut : signalons notamment univoque, d’interprétation simple, conceptuellement identifiée, structurellement définie. Quant au troisième aspect, nous y incluons la mesure dite de consistance interne (p. ex. le coefficient α de Cronbach, 1951 ou le ω de McDonald, 1985), la certification d’une version traduite à partir d’un test original validé dans une autre langue (ou culture), la fidélité (en mode test-retest).

À la lumière de cette analyse et de ce que la littérature scientifique nous rapporte, il est intéressant de constater sommairement que certaines connotations du mot “valide”, utilisé dans l’énoncé “Cet instrument de mesure est valide”, s’y font rares. Ainsi, très peu de publications font état de l’aspect sensible de la mesure : on présente sporadiquement des études sur la différence de scores ou de profils entre des sous-populations ou sous-groupes de personnes, mais la réactivité du score consécutive à un changement d’état de la personne, à une intervention, un contexte, n’est presque jamais documentée. Dans le cas des applications cliniques, la composante adéquatement normée est très souvent escamotée ou bâclée, par raison d’économie, les quelques auteurs qui s’en soucient se repliant souvent sur les normes du test original qu’ils ont traduit. Aussi, sauf dans les tests de type diagnostique dont c’est l’objet, les développeurs et psychométriciens se soucient peu du pouvoir de catégorisation (ou capacité discriminante, voir LAURENCELLE, 2014) de leurs mesures, sans lequel l’établissement d’un barème d’appréciation ou de classement est fortement aléatoire. Enfin, à notre connaissance, il n’existe à peu près rien sur l’aspect à “conclusion prudente voire responsable”, lequel réfère à la fois au traitement explicite et concrètement articulé des erreurs de mesure associées au scoring (LAURENCELLE, 2016, section F) et au nuancement interprétatif du score attribué, pour les cas potentiellement fréquents où la mesure n’a pas l’univocité voulue, c.-à-d. dont l’interprétation et la portée peuvent différer d’un individu à l’autre ou d’un contexte à l’autre. Notre objet sera maintenant d’aborder le travail de la validation psychométrique et ses procédures.

Une nomenclature des démonstrations de validité

Le propos de l’article portant principalement sur les procédures de démonstration de validité plutôt que sur le concept éluif de validité lui-même, il sied d’établir une nomenclature qui structurera l’exposé. D’entrée de jeu,



le terme “validation” sera utilisé comme équivalent de méthode ou procédure de preuve de la validité, et chaque validation réalisera, efficacement ou non, la preuve d’un aspect de la validité du test concerné.

Trois aspects de validité, plus un. Dans ce contexte lexical, nous distinguons heuristiquement trois aspects majeurs de la validité d’un test, d’un instrument de mesure : l’aspect structurel, l’aspect conceptuel, l’aspect normatif, chaque aspect pouvant s’exprimer en plus d’une preuve. En sus, un mot sur la preuve strictement prédictive s’impose.

L’aspect structurel

- La preuve d’ordre structurel consiste à démontrer que le test présente la ou les dimensions statistiques ou la structure interne correspondant à sa définition conceptuelle. Cette exigence s’applique particulièrement aux tests constitués de plusieurs items, tels un questionnaire ou un test d’aptitude.

L’aspect conceptuel

- La preuve d’ordre conceptuel se rapporte quant à elle à la signification, à l’interprétation (sémantique) à donner aux résultats produits par l’application du test. Cette preuve peut être menée de trois manières différentes, indiquant trois nouveaux aspects (ou sous-aspects) de la validité : l’aspect nominal, l’aspect sémantique, l’aspect critérié.
- La preuve de type nominal se fait par la constatation d’un rapport d’adéquation évidente (au premier degré) entre les contenus du test et le concept métrologique visé. On la désigne par l’expression “validité manifeste”, occasionnellement par “validité échantillonnale”¹ lorsque la preuve procède par une démonstration de représentativité des ingrédients du test par rapport à l’ensemble des composants ou aspects définissant le domaine de la grandeur mesurée. Ce type de validation est systématiquement utilisé dans les tests dits de rendement, par exemple en éducativité, dans plusieurs tests d’attitudes ou de tendances (anxiété, suicide, racisme, etc.), et elle fait partie des outils méthodologiques usuels en recherche qualitative (catégorisation conceptualisante, thématisation, etc.; voir FISCHER, 2006; PROULX, 2019; YARDLEY, 2008).

- La preuve de type sémantique, que nous proposons ici, consiste à identifier le sémantème du test, c.-à-d. le faisceau de significations qu’il transmet : c’est essentiellement ce que désigne le terme “construit” de CRONBACH et MEEHL (1955). L’information produite par le test se présente essentiellement sous la forme de la variance des scores. C’est par elle que la recherche, l’identification et le catalogage des sources de cette variance permettent de caractériser à l’aveugle le concept sous-tendu par le test, ce indépendamment de son possible contenu nominal. Les auteurs parlent aussi de validité nomologique ou conceptuelle². Autrement dit, cette preuve fait pendant à ce qu’on appelle traditionnellement “validité convergente” globale, en ayant recours à toutes les sources (tests, mesures) qui corréleront avec le test à valider et dont la somme intégrée des interprétations fournit le “construit” du test à l’étude.
- La preuve de type critérié, l’une des premières proposées pour la validation de tests psychométriques (GUILFORD, 1946; ANGOFF, 1988), consiste à démontrer le lien substantiel et à corrélation désatténuée élevée entre le test et une mesure critère. Le critère utilisé est soit une mesure indépendante, équivalente et ayant idéalement un format différent par rapport au test étudié (validité concomitante ou concurrente), soit une mesure d’une caractéristique fonctionnellement associée au test à titre de cause ou d’effet (validité prédictive).³

Pour mémoire, une corrélation désatténuée est la valeur du coefficient de corrélation corrigée à la hausse pour compenser la non-fidélité d’un ou de ses deux ingrédients. P. ex., la corrélation $r(X, Y)$ désatténuée pour la variable X serait $r_d(X, Y) = r(X, Y) / \sqrt{r_{XX}}$, où r_{XX} est le coefficient de fidélité de X .

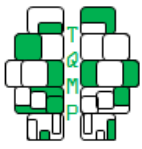
L’aspect normatif

- La preuve normative, rattachée à l’interprétation normative des résultats du test, se fait en établissant ou en vérifiant les normes populationnelles appropriées à la clientèle de répondants visée par le test, notamment pour ses applications cliniques. Sommairement, il s’agit d’établir ou de vérifier le barème interprétatif selon le degré d’importance ou d’exceptionnalité des

1. La preuve de validité échantillonnale (LAURENCELLE, 1998) consiste à montrer que, dans les éléments basant la mesure, on retrouve les caractéristiques et composantes de la grandeur qu’on veut quantifier. Or, dans certains cas, cette grandeur peut être hétéroclite et peu définissable, p. ex. “l’habileté au basket-ball”, “l’employabilité pour une fonction”, “la compétence dans les opérations d’arithmétique enseignées au programme”. La preuve pour ces cas s’opère par un simple calcul de représentativité de l’échantillon des composantes ou items du test ou de l’épreuve. CRONBACH (1971) (p. 457, ‘Irrelevance of correlational evidence’) réfère à ce type de preuve, arguant que les items constituant un test doivent d’abord représenter adéquatement le concept à évaluer, leurs corrélations mutuelles n’étant pas requises.

2. L’épithète “conceptuelle” est classiquement attribuée au concept manifeste global qui émane de l’ensemble des items d’un groupe et que l’analyse corrélationnelle ou factorielle confirmerait. Dans notre nomenclature, cette ‘évidence conceptuelle’ tombe plutôt dans la catégorie de la preuve nominale en raison de son caractère a priori.

3. Dans le cas d’un critère en concomitance, et puisque les deux mesures confrontées reflètent censément le même concept, la corrélation attendue doit jouxter 1 après désatténuation. Dans le cas prédictif, la place respective des deux mesures dans la chaîne causale doit être explicitée et mise en balance avec la force de la corrélation, une position proximale des deux mesures commandant une corrélation jouxtant 1 après désatténuation.



scores, c.-à-d. son étalonnage eu égard à la population visée : la justesse de la norme d'une population à l'autre est particulièrement à surveiller, deux populations pouvant être décrites par le même caractère mais le posséder à des degrés différents. Cette preuve, indispensable mais souvent négligée ou galvaudée de l'élaboration d'un test, se situe hors du propos de cet article : nous référons le lecteur intéressé aux sources suivantes : KANE (2006) et LAURENCELLE (2016).

L'aspect prédictif ou empirique

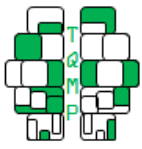
- En ajout, la preuve empirique ou d'ordre strictement prédictif mérite ici une mention à part, en sus de l'argument prédictif déjà invoqué en lien avec la validation conceptuelle de type critérié. Il existe en effet des grilles prédictives, collections d'indicateurs, d'items, de mesures, qui peuvent n'avoir entre eux rien de commun ni de significations globales autres que de contribuer ensemble à 'prédire' un critère déterminé : la seule ou la principale 'validité' de ces grilles est leur utilité pratique (LAURENCELLE, 1998), avec ou sans rapport de causalité explicite. Nommons dans cette optique les grilles médicales diagnostiques, les échelles du MMPI (Minnesota Multiphasic Personality Inventory, versions 1 et 2 : voir HATHAWAY et MCKINLEY (1943) pour la validation par la méthode dite "des groupes contrastés"), les listes d'indicateurs récoltés permettant de circonscrire une compétence pratique, déterminer une habilitation. À notre connaissance, STRONG (1927) serait le premier à avoir expérimenté cette approche, élaborant un test d'intérêts vocationnels à partir des profils de réponses chez diverses catégories de travailleurs professionnels à des fins d'orientation professionnelle et de sélection. Le statut particulier de cette preuve dans le contexte indiqué tient à ce que ces instruments de mesure ne constituent pas à proprement parler des échelles. En effet, ils n'ont pas forcément de métrique interprétable ni de consistance interne ou d'homogénéité sémantique : leurs ingrédients n'y sont réunis que parce que leur participation conjointe permet la prédiction, l'identification, le classement d'un objet donné, de sorte qu'on pourrait invoquer ici la notion paradoxale de "consistance externe" du test. PARISIEN (2021) retrace l'évolution du célèbre inventaire MMPI et ses différents avatars. Dans ce contexte, il analyse par le détail la différence entre l'approche empirique contrastante et l'approche classique corrélationnelle et il en compare les mérites. Les instruments de type prédictif ou empirique ne présentent essentiellement ni structure particulière ni sémantème identifiable, seule leur capacité prédictive en détermine la valeur. Ainsi, ni l'information portée par l'alpha de Cronbach ni celles des analyses facto-

rielles n'y sont pertinentes. L'exposé portant ici sur la validation des échelles psychométriques, nous ne reviendrons pas sur ce sujet. Le lecteur intéressé peut consulter BEAUMIER (2019) pour des références et une discussion articulée de la nature et de l'utilité des grilles prédictives dans les applications en santé. Un sous-ensemble de la catégorie structurelle désigné "test à facettes" (voir "Anneau simple" plus loin) se situe dans le voisinage méthodologique de la preuve par prédiction empirique.

Quatre catégories de la structure d'un test. Couplées à ces trois principaux aspects de la validité, nous posons quatre catégories principales de la structure d'un test : le test à concept simple, le test à facettes, le test complexe ou la batterie, le modèle comportemental.

L'argument de structure repose sur le concept de dimensionnalité entendu dans son acception générale, c.-à-d. se référant au nombre de dimensions de sens ou de mesure que comporte le test. En psychométrie, la préoccupation sur la dimensionnalité conceptuelle déborde souvent, comme ici, sur sa réalisation statistique : c'est le cas pour les différentes approches de l'analyse factorielle, soit principalement l'analyse en axes factoriels principaux (AFP), l'analyse en composantes principales (ACP), l'analyse factorielle confirmatoire (AFC); la modélisation en équations structurelles (MES), et pour la théorie des réponses aux items (TRI). Comme on verra bientôt, la propriété d'unidimensionnalité constitue un enjeu critique dans la validation d'un test : le score obtenu par le répondant à un test doit avoir 'un sens', selon les deux principales acceptions de l'expression : 'sens' parce que sa valeur doit se rapporter à la grandeur d'une qualité, habileté, attitude du répondant, et 'un' parce que ce rapport ne doit pas être ambigu ni contaminé par une ou plusieurs autres qualités du même répondant : le score produit par le test doit varier sur un seul axe. Néanmoins, vu la complexité inhérente de ce qui est estimé chez l'humain et les nombreuses interactions internes pouvant affecter la qualité visée chez le répondant, l'unidimensionnalité stricte est utopique, et nous endossons la proposition de STOUT (1987; voir aussi HATTIE, 1985, 1996) à l'effet d'admettre une "essential unidimensionality", une unidimensionnalité pratique, ou dominante, tout en restant scrupuleux sur son traitement.

- Le test à concept simple. Basé généralement sur une collection d'items (questions, problèmes, essais), le test à concept simple est par définition uniconceptuel, c.-à-d. que son score exprime une seule grandeur métrologique (l'introversion, la capacité aérobie, la xénophobie, l'aptitude pour l'addition à deux chiffres), et idéalement unifactoriel, c.-à-d. qu'un seul groupement statistique des données en rend toute la variance utile.



- Le test à facettes. Ce test, uniconceptuel lui aussi, est ici foncièrement multidimensionnel et peut être plurifactoriel, chacune des facettes formant une structure auxiliaire rapportée au concept central mais distincte de lui (par exemple ‘l’estime de soi’). La structure statistique du test à facettes peut être diverse (anneau simple ou lié, anneau avec noyau, etc.), chaque facette exprimée par un groupe d’items étant constituée d’un ou de quelques éléments (ou sèmes, en linguistique) du concept central (ou sémantème). Nous revenons plus loin sur la structure et les particularités d’approche du test à facettes.
- Le test complexe, la batterie. Certains tests, tels les échelles de développement ou les tests d’attitudes tout comme les “batteries”, comportent deux ou plusieurs échelles, produisant chacune un score qui peut être indépendamment interprété. Chaque échelle, avec le score qui en résulte, doit être validée à titre individuel, selon les prescriptions appropriées aux tests simples ou à facettes : la démonstration de validité est donc plurielle, chaque échelle devant être séparément traitée. Dans plusieurs cas cependant, un score global est aussi proposé, étant obtenu grâce à une combinaison des différents scores des échelles contributives. La validation de ce score global, qui est à la fois pluriconceptuel et plurifactoriel, fait appel à une démarche de second ordre, démarche basée sur une conception globale du domaine de mesure défini par les ingrédients qui le composent et rapportant le score global à une grandeur théorique correspondante, c’est-à-dire logiquement liée au concept global concerné. Différentes échelles de développement peuvent ainsi confluer vers une mesure globale de maturité, les échelles d’attitudes ou d’aptitudes en vue d’une mesure d’employabilité, les habiletés cognitives sommées à titre d’aptitude intellectuelle (ou QI), des indicateurs héréditaires, nosologiques ou comportementaux vers un niveau de risque pathologique, ainsi de suite. À défaut d’une théorie solide et démontrée de l’interaction positive entre les dimensions et échelles à combiner, ledit score global relève du cadre de la grille prédictive, évoqué plus haut.
- Le modèle comportemental. Par modèle comportemental, nous faisons référence aux structures théoriques en réseau proposées par certains auteurs et censées représenter en détail le processus dynamique et déterministe d’un comportement ciblé ou d’un changement de comportement. Le modèle de AJZEN (1991), la Théorie du comportement planifié, est l’un des plus connus. Dans ce modèle, le comportement étudié se voit être la résultante d’une intention d’agir et du contrôle perçu, lesquels subissent l’influence des

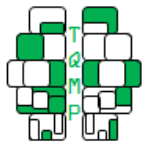
normes subjectives et de l’attitude du sujet par rapport au dit comportement. Des relations directionnelles ou bidirectionnelles relient entre elles les articulations du modèle comportemental, chacune constituant un concept mesurable, soit un état ou une action directement observable, le score d’une échelle de mesure ou bien un concept-fantôme à grandeur variable, ce dernier étant désigné “variable latente” en psychométrie classique. Tel que décrit, le modèle comportemental comprend des articulations, c.-à-d. des composantes mesurables, que relient entre elles des relations de conjonction, d’influence ou de causalité. Chaque composante, avec les mesures qui lui sont associées, mérite sa propre validation, tout comme chaque relation doit être vérifiée, si possible par une preuve expérimentale rarement faite (TRAFIMOW & EARP, 2016; ANDRÉ et collègues, 2015; ANDRÉ & LAURENCELLE, 2020) : l’établissement d’une simple corrélation n’accrédite pas l’existence d’une relation d’influence entre les composantes corrélées, que cette influence supposée soit uni- ou bidirectionnelle.

On doit donc et on peut “valider” chaque instrument de mesure individuel, qu’il soit défini comme simple ou à facettes. Quant aux structures plus complexes (batteries, inventaires, QI, modèles comportementaux), si leurs mesures-ingrédients sont à valider séparément, la structure elle-même doit l’être en tant que telle, globalement, avec toutes ses interactions, ce en la soumettant à des preuves expérimentales sérieuses.

Valider un instrument de mesure simple

Valider un instrument de mesure simple, une échelle de mesure, est sans doute l’assignation de validation psychométrique la plus courante. Il s’agit d’un test qui mesure prétendument un concept simple, c.-à-d. une grandeur à définition sans nuances, monochrome. C’est le cas par exemple des tests de rendement en arithmétique ou en vocabulaire, de certains tests d’attitude, des épreuves d’habileté mentale ou psychomotrice, et cætera.

La preuve structurelle consiste à montrer qu’il ne se trouve qu’une seule dimension de mesure dans le test, vérifiable par un argument statistique ou nominal. La démonstration d’existence d’un seul facteur, idéalement au moyen de l’AFP (voir plus loin), donne la preuve statistique. Pour la preuve nominale conceptuelle, il s’agit de montrer qu’une variation de la mesure correspond à une variation cohérente de la grandeur visée, toutes choses étant égales par ailleurs, ou que le contenu expertisé du test, par exemple un questionnaire ou un test de format verbal, reflète avec évidence la grandeur visée et rien qu’elle.



Preuve structurelle : on peut démontrer l'existence d'une seule dimension par...

- l'examen du contenu. L'homogénéité formelle des items et l'unité sémantique apparente du contenu du test, avérées par la méthode des juges, peuvent suffire.
- le coefficient α , souvent invoqué et de valeur probante douteuse (CORTINA, 1993; STREINER, 2003; SIJTSMA, 2009; LAURENCELLE, 2021, voir note à la section Fidélité, plus bas). L'interprétation d'un α inférieur à son maximum 1 reste équivoque, ce résultat pouvant recouvrir deux ou plusieurs groupements ou dimensions d'items. Quant à une valeur α proche de 1, elle n'est pas rédhitoire mais accuse surtout de la redondance, voire de la répétitivité, des items entre eux.
- l'analyse en axes factoriels principaux (AFP), voire une forme de l'analyse factorielle confirmatoire (AFC),⁴ peuvent servir à démontrer qu'il y a un et un seul facteur présent et que ce facteur reflète une majorité écrasante de la variance, la variance résiduelle étant poussière.⁵ L'analyse parallèle (PA), comme aussi l'approche NEST plus raffinée de ACHIM (2017), notamment ses modalités NESTip ou NESTif, sont à recommander : voir aussi SLOCUM et ZUMBO (2011). Faut-il rappeler que la preuve d'unidimensionnalité doit être appliquée au score (brut) d'échelle, non pas à son score factoriel, lequel est de facto unidimensionnel. Noter que, ici comme ailleurs pour la validation d'une échelle psychométrique, l'application de l'AFC avec aménagements structurels est aporétique et à proscrire.
- d'autres évaluations de l'unidimensionnalité statistique, par exemple celles de l'arsenal de la Théorie des réponses aux items (TRI) : voir BERTRAND et BLAIS (2004), p. 202 et suivantes, et STOUT (1990). L'indice ω_h de McDonald (1985; voir aussi Zinbarg et coll., 2005) s'applique aussi, voire le simple pourcentage de variance (vraie) ou de valeur propre associée au premier facteur primaire (voir %F1, plus bas).
- la corrélation proche de 1 avec une mesure-critère sûre, après désatténuation pour la mesure critère, ce qui assurera à la fois la preuve structurelle et la preuve sémantique.

Preuve conceptuelle nominale : on peut vérifier la nature de la grandeur mesurée par :

- l'examen du contenu par évidence justifiée : la littérature parle alors de validité manifeste (*face validity*) ou de validité échantillonnale (il s'agit alors d'un échantillonnage homogène⁶ des contenus). Arguer sur l'évidence objective (pertinence et homogénéité) des items par rapport à la grandeur à estimer, leur évidence lexicale s'il s'agit d'items verbaux, ou matérielle s'il s'agit d'éléments ou de situations concrets (adéquation et similitude des éléments, situations ou série d'essais analogues, etc.).

Preuve conceptuelle sémantique : on peut aussi vérifier la nature de la grandeur mesurée par...

- une corrélation proche de 1 ou une correspondance presque parfaite avec un critère sûr (après désatténuation pour le critère), laquelle assure à la fois la preuve structurelle et la preuve sémantique. Les auteurs parlent ici de validité concomitante (ou *concurrent validity*).
- l'identification du sémantème du test par son positionnement dans une "toile nomologique" (CRONBACH & MEEHL, 1955) ou réseau de corrélations signifiantes approprié, par un examen différentiel sur un tableau de relations multitraits-multiméthodes (CAMPBELL & FISKE, 1959) ou par son statut corrélationnel probant dans un modèle en AFC à khi-deux non-significatif.

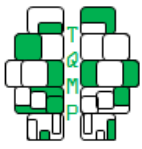
Preuve conceptuelle critériée : on peut aussi vérifier la nature de la grandeur mesurée par :

- une corrélation élevée ou une correspondance significative avec un critère reconnu (après désatténuation pour le critère). La preuve sémantique est établie en avérant que la mesure exprime un effet observé d'une variation-critère contrôlée, ou qu'elle quantifie la cause d'un effet-critère contrôlé. Les relations cause-effet mentionnées doivent naturellement être établies préalablement et reconnues. Les auteurs parlent ici de validité prédictive.

4. La distinction entre dimension conceptuelle et dimension structurelle d'un test psychométrique est notamment signalée par BERTRAND et BLAIS (2004, p. 202 et suivantes).

5. Dans plusieurs cas de figure, l'examen factoriel fait apparaître un premier facteur dominant, la variance résiduelle conséquente n'atteignant pas la force voulue ni l'évidence recherchée pour former un ou quelques autres facteurs. Cette situation montre que le premier facteur et son expression numérique renvoient l'essentiel de l'information présente. Cependant, il se peut qu'après la déduction de ce facteur la variance résiduelle reste tout de même significative (évaluable par un test du khi-deux), de sorte que le score brut lui-même, exprimant surtout l'information du premier facteur, reste contaminé par cette variance et biaisé par rapport au concept visé. Un test simple de ce risque consiste à faire la différence entre la fidélité (via le coefficient $r_{XX'}$) du score brut et la proportion de variance attachée au facteur retenu, la différence, soit la part de variance spécifique relative du score brut, devant avoisiner zéro.

6. Les différentes composantes constituant le groupe sémantique du concept doivent être homogènes, c.-à-d. constituer des avatars interchangeables (même domaine sémantique, même mode de mesure) d'un même modèle conceptuel. Quant aux facettes, dans le cas du concept à facettes, elles doivent présenter une dose d'hétérogénéité (théoriquement convergente) qui assure leur existence individuelle distincte.



Quelques arguments complémentaires pour la validation d'un test psychométrique

Examinons maintenant, sommairement, quelques méthodes et techniques permettant d'argumenter en faveur d'autres aspects de la validité d'un test.

La fidélité : montrer que la mesure permet de quantifier solidement une grandeur provisoirement stable chez la personne évaluée et par le fait même de classer un même groupe de personnes de façon consistante d'une occasion à l'autre; l'indice privilégié est le coefficient de fidélité r_{XX} , dit du test-retest. La précision ou stabilité du score est quantifiée via l'erreur-type de mesure (σ_e), obtenable par la formule définitionnelle $\sigma_e = \sigma_X \cdot \sqrt{1 - r_{XX}}$ (voir cependant plus bas pour l'estimation de σ_e) et la consistance de classement par l'indice $CC = 1/2 + (\sin^{-1} r_{XX})/\pi$ en radians (LAURENCELLE, 1993, 1994). La procédure classique du test-retest, la seule valable à notre connaissance, revient à mesurer deux fois, en deux circonstances indépendantes sous des conditions équivalentes, un échantillon raisonnablement représentatif d'individus de la population concernée et d'en calculer la corrélation $r(X_1, X_2)$, notée r_{XX} . En théorie des tests, cette quantité dénote la proportion de variance vraie de la mesure, la proportion complémentaire, $1 - r_{XX}$, étant attribuée à l'erreur de mesure.

Dans le cas de tests uni-conceptuels formés de deux ou plusieurs items, certains auteurs proposent d'autres techniques ne requérant qu'un seul moment de mesure et basées sur le morcellement du test (corrélation entre deux moitiés du test, technique de Rulon, etc. : voir Laurencelle 1998 ou Bertrand et Blais 2004), les plus connues de ces techniques étant le coefficient α (alpha) de CRONBACH (1951) (par émiettement du test en ses éléments individuels) et le ω (oméga) de McDONALD (1985) (basé sur l'analyse factorielle). La légitimité (pour ne pas dire, la validité) de ces techniques pour estimer la fidélité, notamment la stabilité de la mesure d'une occasion à l'autre et la portion de variance vraie des scores, dépend de conditions étrangères à l'effectuation de la technique, conditions qui sont souvent difficiles à satisfaire et la plupart du temps ignorées. L'argument suivant suffit à disqualifier le α et la plupart des techniques cousines comme estimateurs de fidélité.

Soit théoriquement k grandeurs mutuellement indépendantes qui caractérisent chaque objet, par exemple une personne, parmi un ensemble N d'objets. Pour chaque objet donné, les grandeurs resteraient de valeur stable, tout comme l'agrégat des k valeurs, répétable d'une mesure à l'autre : le score obtenu en totalisant ces k valeurs serait quasi parfaitement fidèle (ou $r_{XX} \approx 1$) d'une occasion à l'autre, et sa "consistance interne" nulle

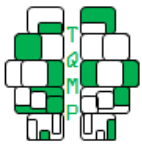
($\alpha, \omega \approx 0$) par définition puisque basée sur des intercorrélations nulles des grandeurs. D'autre part, soit théoriquement k grandeurs mutuellement apparentées (à degrés divers) et intercorrélées (selon $\alpha, \omega > 0$), grandeurs qui caractérisent l'état momentané de chaque objet ou personne, parmi un ensemble d'objets ou de personnes. Pour chaque objet donné, les grandeurs fluctueraient selon les conditions du contexte régnant au moment de la mesure, de sorte que le score total produit pour chacun fluctuerait d'un moment à l'autre, selon le contexte où il se trouve. Pour le score résultant, la consistance interne (p. ex. α ou ω) à chaque moment de mesure resterait positive et peut-être élevée (selon le degré d'affinité des grandeurs mesurées), alors que la fidélité test-retest reculerait sérieusement ($r_{XX} < 1$ et $r_{XX} \rightarrow 0$).

Les coefficients α et ω , comme les autres indices issus d'un seul moment de mesure, ne sont donc pas des estimateurs sûrs de la fidélité : leur valeur probante dépend essentiellement de la dimensionnalité de l'ensemble des éléments que le score qualifié regroupe et la reflète (voir aussi LAURENCELLE, 2021). Par contraste, l'indice de fidélité r_{XX} par test-retest atteste seulement et sûrement qu'il y a dans le score obtenu une caractérisation solide de l'objet mesuré, quelle que soit la dimensionnalité des éléments qui le composent.

La sensibilité : montrer que, pour la même personne ou le même groupe de personnes, la mesure reflète la variation de grandeur due à un changement signalé des conditions pertinentes où les personnes se trouvent ou à une intervention externe ciblée sur cette grandeur. Nonobstant sa fidélité, ou stabilité pour des conditions de mesure comparables, le test doit pouvoir refléter les variations d'état de la personne lors d'un changement réel de conditions.

Soit X_0 une mesure de référence prise sur un individu, σ_X l'écart-type (ou son estimateur s_X) des mesures de la population concernée, X_C la mesure après intervention et γ , un seuil de significativité statistique ($0 < \gamma < 1$) : le test $(X_C - X_0) \geq \sigma_X \cdot \sqrt{1 - r_{XX}} \cdot \sqrt{2} z_{1-\gamma}$, où $z_{1-\gamma}$ est le centile $100 \cdot (1 - \gamma)$ d'une variable normale standard, indiquera que le changement positif obtenu chez un individu donné peut être considéré notoire ou réel, en admettant un risque d'erreur de $100\gamma\%$: dans le cas de variation négative, il suffit d'inverser l'inégalité ci-dessus. Pour $\gamma = 0,05$ (incertitude unidirectionnelle) ou $0,025$ (incertitude bidirectionnelle), la quantité à battre est $\sigma_X \sqrt{1 - r_{XX}} \times 2,33$ ou $\sigma_X \sqrt{1 - r_{XX}} \times 2,77$ respectivement.

Le pouvoir de catégorisation et la précision : L'aptitude d'un test, d'une mesure, à bien différencier et catégoriser les valeurs de la grandeur mesurée dépend en bonne partie de la précision ou finesse de sa quantification. Cette précision peut déjà être caractérisée par l'erreur-type de mesure (σ_e) définie plus haut : le score produit s'écartera



plus ou moins de la valeur juste, selon l'importance de l'erreur-type, laquelle covarie aussi avec l'unité de mesure appliquée. Cet écart brouille les différences de valeurs entre objets numériquement voisins. ALLAIRE et LAURENCELLE (1998) proposent deux estimateurs excellents de cette quantité σ_e découlant de la procédure test-retest, soit leurs formules notées V3 : $\sqrt{s_1 s_2 (1 - r_{X_1, X_2})}$ ou V5 : $s(X_1 - X_2) / \sqrt{2}$, respectivement, où r_{X_1, X_2} est la corrélation des deux mesures en test et retest et $s(X_1 - X_2)$ dénote l'écart-type des différences de scores individuelles entre les deux.

Un test à erreur-type plus petite pourra donc attribuer une mesure, c.-à-d. une catégorie numérique particulière, mieux définie, plus étroite pour ainsi dire, qu'un test à large étendue d'erreur. La capacité discriminante quantifie ce pouvoir de catégorisation d'une mesure, en indiquant le nombre approximatif d'intervalles statistiquement distincts qu'un test, une mesure, peut établir. S'inspirant de BLOOM (1942) et du "pouvoir classificatoire" de FERGUSON (1949), LAURENCELLE (1997, 2014) met au point le concept opérationnel de capacité discriminante (D_ρ) et en propose deux quantifications, soit $D_\rho = 2,67 / \sqrt{1 - r_{XX}}$ ou $2\sqrt{(1 + r_{XX}) / (1 - r_{XX})}$, ρ_{XX} ou son estimateur r_{XX} étant encore l'indice de fidélité de la mesure.

La justesse normative : Des mesures de tout domaine, comme celles du poids des aliments en épicerie, de la taille d'un jeune enfant, de la glycémie d'un patient ou le score obtenu à un test d'aptitude mécanique ou de tendance suicidaire, servent à éclairer sur le prix à payer, documenter le diagnostic à faire, décider d'intervenir ou non pour l'emploi sollicité ou la santé d'une personne. L'instrument, le test qui fournit ces mesures doit évidemment être fidèle (précis) et valide, mais, pour cette fin d'interprétation comparative et de décision, il doit aussi être calibré, c.-à-d. que son barème d'interprétation doit être ajusté et appuyé sur des critères de réalité, donc étalonné. Le premier de ces critères, dans le cas des tests psychométriques, repose sur la qualité des normes (qui basent le barème), dont la justesse (le non-biais moyen) et la capacité de représentation (l'étendue suffisante) de la population visée sont les premières qualités (AERA, 2014; KANE, 2006; LAURENCELLE, 2016). Dans le cas des tests à portée diagnostique ou classificatoire, le barème doit considérer en outre les sous-populations de référence (p. ex. sexe, âge, groupes pathologiques ou échantillons de travailleurs ou de personnes qualifiées servant de références, race).

La prudence d'attribution : Bien qu'on ait établi sa fidélité et un ou quelques aspects de sa validité, toute mesure, et la mesure psychométrique en particulier, garde souvent une part d'incertitude numérique non négligeable, et toute déclaration, décision ou classement qui en sont tirés en seront affectés : refus injustifié de recrutement pour un emploi, mauvais diagnostic médical ou psychologique, atteinte notée insuffisante d'un critère de performance, etc. La disponibilité et la prise en compte explicite de l'erreur de mesure (voir Le pouvoir de catégorisation et la précision, plus haut) devraient être obligatoires dans ces opérations, et l'on constate hélas qu'elles en sont presque partout absentes. C'est dans ce contexte de classement et de décision que LAURENCELLE (2015, 2016, 2017) propose son concept de "norme sûre", basé sur un modèle mathématique intégré des erreurs impliquées (l'erreur de mesure et l'incertitude échantillonnale des quantiles) pour la catégorisation et la sélection prudentes et légalement responsables des répondants.

La valeur utilitaire : Pour qu'il y ait quelque intérêt à s'en servir, un test doit certes être utile, en plus de ses autres qualités déjà documentées. Utile à qui, utile à quoi, nous demandera-t-on. Nous référons ici à l'utilité externe du test, son application à et pour des personnes, dans des contextes de traitement, de sélection, de promotion, d'expertise.⁷ Or, pour certains tests, l'utilité prédictive est la seule propriété, la seule qualité essentielle : c'est le cas des grilles prédictives (KUHN & JOHNSON, 2013; STEYERBERG, 2010; voir BEAUMIER, 2019, pour un exposé nuancé sur cette question). La grille prédictive, qui prend souvent la forme d'une liste de vérification (ou *checklist*), est constituée d'un groupe d'items (grandeur mesurée, présence/absence d'une caractéristique, échelles Likert, données biométriques ou démographiques, etc.), groupe de caractéristiques mis en regard d'un critère externe à prédire (grandeur complexe ou coûteuse à mesurer directement, degré de gravité d'un état, niveau d'employabilité d'un candidat, etc.) et soumise à une validation empirique (voir plus haut, "L'aspect prédictif et empirique"). Les items prédictifs ainsi groupés n'ont pas nécessairement de liens l'un avec l'autre et, idéalement, ils n'en ont pas, leur fonction étant de contribuer chacun à prédire un aspect ou une portion spécifique du critère concerné.⁸ L'argument de validité échantillonnale s'applique ici vis-à-vis d'une population d'items à caractéristiques hétérogènes : le groupe d'items ainsi constitué n'a pas de concept nucléaire de référence, de validité manifeste ni de consistance interne

7. Les tests, on le sait, peuvent aussi être et sont souvent "utiles" sans vraiment être exploités pour des applications pratiques, cliniques ou autres. Les chercheurs psychométriciens développent des tests et analysent les données produites pour comprendre et modéliser les qualités humaines visées par ces tests, cela dans le but d'en élaborer et vérifier une théorie. L'utilité pratique du test n'est pas alors considérée, voire pertinente.

8. En fait, une grille idéale serait composée d'items tous mutuellement indépendants (sans intercorrélation), non seulement pour en mieux asseoir l'efficacité prédictive mais aussi pour protéger la fonction prédictive contre des effets compensatoires éventuels (signalés par l'apparition de coefficients prédictifs de signe opposé au coefficient de corrélation direct de l'item), effets susceptibles d'introduire de la distorsion dans le calcul.

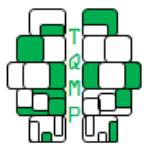
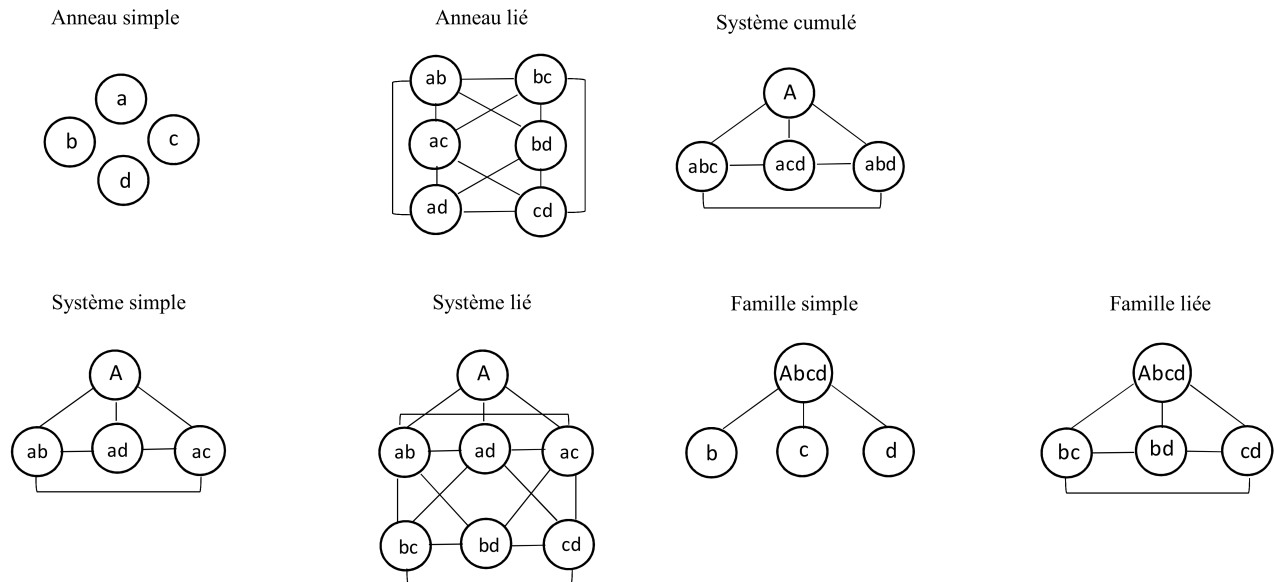


FIGURE 1 ■ Diagrammes présentant les structures possibles des concepts à facettes



intéressante. Une théorie mathématique complète de la validation d'une grille prédictive reste à construire.

Valider un instrument de mesure à facettes

Qu'est-ce qu'un concept ou un test à facettes? Louis Guttman, dans son article "An outline of some new methodology for social research" paru en 1954, a consacré la notion de facette et, par le fait même, celle du concept à facettes, en donnant la définition suivante (p. 399) :

Considérons un ensemble A d'éléments $a_1, a_2, \dots$, et un ensemble B d'éléments $b_1, b_2, \dots$. Soit C, le produit de A et B : $C = A \times B$. Un élément typique de C, disons c , est une paire d'éléments $c = (a_j, b_k)$, l'un provenant de A, l'autre de B. Si A possède m éléments et B en a n , alors C possède $m \times n$ éléments. Nous dirons que C est un ensemble à deux facettes, et que A et B en sont les facettes. En général, C peut être le produit d'un nombre quelconque de facettes, pas seulement deux. (traduction libre)

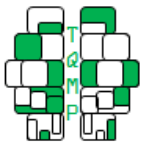
Mettant de côté la rigueur mathématique un peu scolaire de GUTTMAN (1954), CARVER (1989) présente une définition généralement mieux reçue (p. 577) :

Par concept à facettes, j'entends un concept composé de deux ou plusieurs concepts subordonnés, chacun pouvant être conceptuellement distingué des autres et mesuré séparément, tout en étant reliés les uns aux

autres aussi bien logiquement qu'empiriquement. (traduction libre)

L'idée de concept à facettes a fait florès chez les chercheurs et dans les modèles explicatifs des sciences humaines, et elle a aussi été évoquée mais plus discrètement en psychométrie. GUTTMAN (1954, voir aussi GUTTMAN et GREENBAUM, 1998), précisément dans une optique psychométrique, relève la difficulté qu'il y a à faire l'analyse statistique du concept à facettes, dénonçant notamment l'inadéquation de l'analyse factorielle pour ce cas. Il propose plutôt l'examen du tableau de corrélations et la technique en partie visuelle dite du *Smallest Space Analysis* (SSA) (GUTTMAN, 1968, 1982; SCHLESINGER & GUTTMAN, 1969; BLOOMBAUM, 1970).

La contribution de GUTTMAN (1954) a été, comme nous l'avons écrit, de consacrer le terme "facette" en en faisant un élément structurel explicite et ordinairement mesurable d'un test psychométrique. Il reste que, dès l'origine de la psychométrie, l'élaboration et la structure des tests psychométriques comportaient souvent un inventaire d'aspects complémentaires de la grandeur à évaluer, une sélection de caractéristiques du concept à cerner. Binet et Simon le font déjà en 1904 pour la mesure de l'intelligence et d'autres s'y sont depuis exercés pour celle des attitudes, des dispositions émotionnelles, des intérêts dits vocationnels, etc. Cette ubiquité d'une structure en blocs sous-jacente à nombre de tests psychométriques, que cette structure soit explicite ou qu'elle y dorme incognito, re-



hausse l'intérêt qu'il y a à investiguer la mécanique du concept à facettes, ses diverses manifestations, son analyse et les assises possibles de sa validation. L'analyse par facettes apparaît vraiment ici comme une perspective de rechange par rapport à l'analyse par 'facteurs', laquelle est traditionnellement assujettie aux approches linéaires et corrélationnelles.

Le concept à facettes est donc une structure psychométrique, et cette structure peut présenter diverses formes (Figure 1). En extension de GUTTMAN (1954), posons qu'une facette est concrétisée par un item, une échelle d'un test, une mesure, et que, à son tour, la facette reflète une ou quelques composantes psychométriques pures de l'objet évalué, lesquelles sont supposément appropriées ou non au concept global. Aux fins d'illustration et sur cette base après, nous avons imaginé quelques structures possibles du concept à facettes, sept au total, que nous définissons ci-après.

1- L'anneau simple (AS) : a, b, c, d

Le concept d'anneau est défini par la réunion de k facettes, que nous présentons ici par un groupement de k cercles, chaque cercle exprimant une composante ou un groupe de composantes pures de l'objet évalué : celles-ci sont représentées ici par les lettres de l'alphabet : l'exemple donné en compte 4, indiquées par les lettres a, b, c et d . Les composantes sont en principe indépendantes (GUTTMAN, 1954), sans l'être nécessairement. Elles sont mutuellement complémentaires et leur réunion (en contenus conceptuels comme en variance) constitue le concept.

2- L'anneau lié (AL) : ab, ac, ad, bc, bd, cd

Les facettes de l'anneau sont ici au moins partiellement liées, du fait qu'elles partagent une ou plusieurs des 4 composantes présentes : chaque lien est représenté ici par une ligne. Dans l'exemple fourni, les 6 facettes ont 12 liens mutuels, en partageant au moins une composante (p. ex. ab vs bd), et 3 facettes sont mutuellement indépendantes (p. ex. ab vs cd), n'en partageant aucune. Noter que la facette ab peut être représentée numériquement par la combinaison (linéaire) $f(ab) = a + b$ (sans pondérations, comme ici, ou avec pondérations). D'autres formes d'agrégation des composantes sont possibles.

3- Le système simple (SS) : A, ab, ac, ad

Le concept est nucléé, possédant une facette centrale ou noyau, représentée ici par A , à laquelle s'ajoutent des facettes satellites. L'exemple fait voir le noyau A (la majuscule connotant le statut de noyau), additionné séparément des composantes satellites associées b, c et d .

4- Le système lié (SL) : $A, ab, ac, ad, bc, bd, cd$

Au noyau A séparément défini, s'ajoutent des facettes complémentaires, ces dernières étant mutuellement bi-relées.

5- Le système cumulé (SC) : A, abc, abd, acd

Au noyau A séparément défini, s'ajoutent des facettes à lui associées et mutuellement bi-relées. La facette abc peut être représentée numériquement par $f(abc) = a + b + c$, sans pondération des composantes (comme ici) ou avec.

6- La famille simple (FS) : A^3bcd, b, c, d

Au noyau A déjà facetté, s'ajoutent trois facettes individuelles simples et mutuellement indépendantes. La notation A^3bcd indique que le poids (en variance) de l'élément a dans le noyau est comparable au poids total des composantes associées. En notation linéaire, cette facette nucléaire serait représentée par $f(A^3bcd) = \sqrt{3} \cdot a + b + c + d$.

7- La famille liée (FL) : A^3bcd, bc, bd, cd

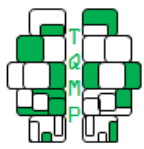
Au noyau A déjà facetté, s'ajoutent les combinaisons mutuellement bi-relées des trois facettes complémentaires.

Nous avons exploré numériquement les modèles esquissés ci-dessus et produit pour chacun un tableau de $N = 5000$ lignes de données créées à partir de composantes indépendantes et identiquement distribuées (i.i.d.) de loi normale standard, notées a, b, c, d . Chaque variable construite a été numériquement ramenée à une variance unitaire, par exemple $a/\sqrt{1}$, $(a + b + c)/\sqrt{3}$, $(\sqrt{3}a + b + c + d)/\sqrt{6}$.

Un examen superficiel des analyses (Tableau 1) faites sur les sept modèles explorés illustre déjà très bien l'assertion de Guttman sur l'inadéquation de l'analyse factorielle classique pour caractériser le concept à facettes.

Les données laissent aussi apparaître la pauvre fiabilité de l'indice dit de factorabilité Kaiser-Meyer-Olkin (KMO) et de l'analyse parallèle (PA) dans ce contexte. Outre sa technique du SSA, que nous n'aborderons pas ici, GUTTMAN (1968) recommande dans chaque cas l'examen 'intelligent', c.-à-d. exploratoire et sans préconception, de la matrice de corrélations : les tableaux reproduits en Appendice fournissent ces matrices de même que les extrants de l'AFP pour tous les facteurs extraits (sans rotation). Noter que les données exploitées ici sont dépourvues d'erreur autre que l'erreur échantillonnale, les corrélations calculées ayant une erreur-type (σ_r) égale ou inférieure à $1/\sqrt{5000 - 1} \approx 0,0141$.

Les preuves de validité d'un test basé sur un concept

**Tableau 1 ■** Indices analytiques des modèles de facettes

Modèle	Composition	AFP	PA	NEST	KMO	%F ₁	%F _k	α
Anneau simple (AS)	a b c d	0	1	0	0,508	1,8	2,2	-0,029
Anneau lié (AL)	ab ac ad bc bd cd	1>3	2	4	0,334	50,1	100	0,801
Système simple (SS)	A ab ac ad	1	1	1	0,757	63,3	65,4	0,862
Système lié (SL)	A ab ac ad bc bd cd	1>1>2	2	4	0,495	48,9	65,4	0,821
Système cumulé (SC)	A abc abd acd	1	1	1	0,831	63,3	65,2	0,869
Famille simple (FS)	A ³ bcd b c d	1	1	1	0,443	38,2	87,7	0,424
Famille liée (FL)	A ³ bcd bc bd cd	1	1	1	0,784	59,8	88,0	0,776

Note. Les calculs pour chaque modèle sont basés sur un ensemble de $N = 5000$ protocoles utilisant des variables normales standard (a, b, c, d) indépendantes. L'analyse factorielle par axes principaux (AFP), l'analyse parallèle (PA) et le NEST (ACHIM, 2017) ont été utilisés. %F₁ dénote le pourcentage de variance vraie associée au premier facteur, %F_k celui de la variance vraie de tous les facteurs extraits. Le coefficient α (alpha) de Cronbach de la matrice de données est aussi fourni.

Chaque composante est représentée par une variable normale standard (de variance 1), sauf la composante A³ affectée d'une variance de 3. Une composante notée ab est produite par la variable $a + b$. Pour la composante A³bcd, elle donne lieu à la variable $(\sqrt{3}a + b + c + d)/\sqrt{6}$.

à facettes, lequel aboutit tout de même à un score principal à interpréter, se trouvent dans l'ensemble ou dans un sous-ensemble des preuves relatives à un concept psychométrique simple, comme nous venons de les parcourir. Les notes suivantes ont pour seul but de faire ressortir certaines particularités de chaque modèle de facetage.

Preuves pour l'anneau simple : Sans noyau central (ni structurel, ni conceptuel), le concept à anneau simple tient son identité et sa nature de la conjonction de ses facettes, assemblées là par décret créatif (à partir d'une définition par collection de caractéristiques) ou empiriquement en fonction d'un critère objectif externe. Dans le premier cas, la validation ressortit donc directement à la preuve nominale, notamment par la démonstration de la représentativité des contenus vis-à-vis de l'objet à évaluer. Cronbach, en 1971, renchérit en ce sens par la boutade : "Les corrélations n'ont rien à voir avec la validation de contenu... Rien dans la logique de cette validation ne requiert que le domaine ni le test aient des contenus homogènes." (ibid., p. 457); ici, même notre coefficient α est nul. Les autres preuves de type conceptuel peuvent convenir aussi. Le second cas, de type empirique, appuie sa validation sur la corrélation individuelle de chaque facette avec un critère ou une mesure objective et externe, tel que dans certaines échelles et indications diagnostiques du MMPI (voir p. ex. PARISIEN, 2021).

Preuves pour l'anneau lié : Le test NEST fait apparaître correctement 4 facteurs distincts, basés sur leurs 4 composantes indépendantes (a, b, c, d) tandis que le facteur 1 (%F₁ = 50, 1) représente bien l'amalgame des facettes. L'analyse du tableau de corrélations permet de reconstituer la structure en 4 composantes distinctes, regroupées en ${}_4C_2 = 6$ couples intercorrélés. La plupart des preuves de validation

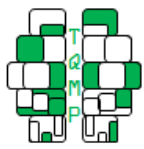
documentées s'appliquent.

Preuves pour le système simple : Il apparaît ici un concept quasi simple (AFP = PA = NEST = 1, %F₁ $\approx$ 63, 3, %F₁ / %F_k $\approx$ 0, 97), à unidimensionnalité dominante, dont la structure facettaire se découvre à la fois dans le tableau de corrélations et la composition du facteur 1. A priori, le poids dominant de la composante a dans le facteur 1 lui signale une importance singulière, le niveau des corrélations entre composantes permet ensuite de découvrir la structure agrégée de ce concept. La plupart des preuves de validation documentées s'appliquent.

Preuves pour le système lié : L'AFP tout comme le NEST reconnaissent 4 facteurs (les 7 variables présentées permettraient en AFP jusqu'à 6 facteurs), qu'une analyse raffinée de la matrice de corrélations rapporte à 4 sources indépendantes, observées en jumelages. Cette analyse logique permet d'abord d'isoler le cœur du concept en l'épinglant à la première variable (A), et sa suite, de découvrir les jumelages des quatre sources impliquées, soit le cœur et trois facettes indépendantes. La plupart des preuves de validation documentées s'appliquent.

Preuves pour le système cumulé : Image analogue à celle associée au système simple (AFP = PA = NEST = 1, %F₁ $\approx$ 63, 3, %F₁ / %F_k $\approx$ 0, 97), hormis que la composition jumelée des trois variables se laisse découvrir grâce à leur niveau spécifique d'intercorrélation. La plupart des preuves de validation documentées s'appliquent.

Preuves pour la famille simple : Basée sur 4 variables, l'AFP (tout comme la procédure NEST) ne peut produire que 3 facteurs eu égard aux 4 composantes présentes. Néanmoins, la structure en facettes ressort très bien du tableau de corrélations. En admettant que la première variable base le concept (le poids au facteur 1 de la première



variable étant plus de deux fois plus élevé que celui des trois autres), ses trois facettes associées y sont clairement jointes, tout en restant mutuellement indépendantes. La plupart des preuves de validation documentées s'appliquent.

Preuves pour la famille liée : Le jumelage séparé des facettes dans ce modèle, comparativement au modèle de famille simple, a pour effet de réduire la dimensionnalité des données, en les exprimant en un seul facteur : pour ce facteur exprimant donc tout le concept, le poids de la première variable, le “noyau”, ne domine pas numériquement le facteur, au contraire du cas Famille simple. Les variances conjuguées des variables satellites contribuent davantage à définir le concept tout en le nuancant. Bien qu'unifactoriel, ce modèle ne présente pas une unidimensionnalité dominante, sa structure diaprée étant renforcée par les jumelages systématiques des composantes (tel qu'apparaissant dans le tableau de corrélations). La plupart des preuves de validation documentées s'appliquent.

Les considérations et conclusions formulées ici sur l'analyse structurelle du concept à facettes sont à la fois optimistes et surtout indicatives : elles sont basées sur des structures de données artificielles et sur des données dépourvues d'erreur de mesure. Quant à l'incertitude échantillonnale touchant le matériau des coefficients de corrélation exploité, son erreur-type d'environ 0,014 est très petite. Dans les cas d'application courants, l'erreur de mesure est présente et souvent substantielle et les tailles échantillonales rencontrées plutôt modestes, dégradant aussi la précision des corrélations calculées : ces deux facteurs rendent alors l'AFP approximative et surtout la configuration du tableau de corrélations plus brouillée.

Conclusion

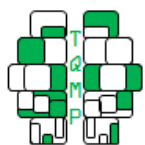
Les propos, exemples et interprétations présentés dans cet essai sont des indications, non des recettes, afin de mener à bien une validation des données d'une échelle de mesure. Notre but premier consistait à faire voir la diversité des contextes et des critères de la validation psychométrique et la nécessité pour l'analyste de reconnaître et prendre en compte cette diversité. Cinq leçons auxiliaires découlent des constats de cette expédition : (1) l'examen minutieux du tableau de corrélations prime sur toute autre analyse; (2) l'analyse factorielle sur axes principaux (AFP) n'est ni universellement utile ni suffisante pour révéler la structure relationnelle des variables; (3) il est instructif et parfois suffisant de considérer la structure factorielle obtenue avant rotations, particulièrement pour l'analyse d'un concept à facettes; (4) il est téméraire d'endosser à l'aveugle l'alpha de Cronbach ou ses avatars sans d'abord mettre au jour la dimensionnalité des données, et

généralement injustifié de l'interpréter comme estimateur de fidélité alors qu'il est par définition un estimateur de la consistance (corrélationnelle) interne de la mesure; (5) dans le cas de données à intercorrélations significatives et à organisation factorielle démontrée, la précision des corrélations, mesurable par leur marge d'incertitude à 95% et plafonnée à environ $4/\sqrt{N}$, doit être considérée.

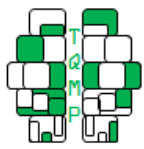
Quant à la validité, elle est essentiellement un jugement porté par quelqu'un, une autorité, une instance décisionnelle, en fonction d'un besoin, d'un contexte et d'une perspective d'utilisation donnés : sera jugée valide la mesure que ladite instance décisionnelle sanctionnera comme satisfaisante. Et, comme ce décret de validité est contextuel et lié à une décision arbitrale, la validité elle-même échappe à toute définition. Il appartient au développeur du test ou à son procureur, le psychométricien, d'en construire et proposer la preuve.

Références

- ACHIM, A. (2017). Testing the number of dimensions in exploratory factor analysis. *The Quantitative Methods for Psychology*, 13, 64-74. doi :[10.20982/tqmp.13.1.p064](https://doi.org/10.20982/tqmp.13.1.p064)
- AERA, A. e. N. (2014). *Standards for Educational and Psychological Testing*. Washington, DC : American Psychological Association et National Council on Measurement in Education.
- AJZEN, I. (1991). The Theory of Planned Behavior. *Organizational Behavior and Human Decision Processes*, 50, 179-211.
- ALLAIRE, D. & LAURENCELLE, L. (1998). Comparaison Monte Carlo de la précision de six estimateurs de la variance d'erreur d'un instrument de mesure. *Lettres Statistiques*, 10, 27-50.
- ANDRÉ, N. & LAURENCELLE, L. (2020). Where do the conceptual models for behavior change come from, and how are they used? A critical and constructive appraisal. *The Quantitative Methods for Psychology*, 16, 1-8. doi :[10.20982/tqmp.16.1.p001](https://doi.org/10.20982/tqmp.16.1.p001)
- 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, 125-148. doi :[10.7202/1036330](https://doi.org/10.7202/1036330)
- ANGOFF, W. H. (1988). Validity : An evolving concept. In H. WAINER & H. I. BRAUN (Éd.), *Test validity* (p. 19-32). Hillsdale, NJ : Routledge.
- BEAUMIER, M. (2019). *Élaboration et validation d'une grille prédictive de la vascularisation artérielle insuffisante à une plaie au membre inférieur*. UQTR-Université de Montréal : Thèse de doctorat (Ph.D.) en sciences biomédicales.



- BERTRAND, R. & BLAIS, J.-G. (2004). *Modèles de mesure*. L'apport de la théorie des réponses aux items. Québec : Presses de l'Université du Québec.
- BINET, A. & SIMON, T. (1904). Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'année psychologique*, 11, 191-244.
- BIPM. (2012). *Vocabulaire international de métrologie – Concepts fondamentaux et généraux et termes associés* (3e édition). Paris : Bureau international des poids et mesures.
- BLOOM, B. A. (1942). Test reliability for what? *Journal of educational psychology*, 33, 517-526.
- BLOOMBAUM, M. (1970). Doing smallest space analysis. *Journal of conflict resolution*, 14, 409-416.
- BORSBOOM, D., MELLENBERGH, G. J. & van HERDEN, J. (2004). The concept of validity. *Psychological Review*, 111, 1061-1071. doi :[10.1037/0033-295X.111.4.1061](https://doi.org/10.1037/0033-295X.111.4.1061)
- CAMPBELL, D. T. & FISKE, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.
- CARVER, C. S. (1989). How should multifaceted personality constructs be tested? Issues illustrated by self-monitoring, attributional style, and hardness. *Personality Processes and Individual Differences*, 56, 577-585.
- CHEVALIER, R., CONTE, R., LABRIE, Y., LANDRY, F. & NADÉAU, A. (1978). *Éléments de métrologie*. Québec : Les éditions La Liberté.
- CORTINA, J. M. (1993). What is coefficient α ? An examination of theory and applications. *Journal of Applied Psychology*, 78, 98-104.
- CRONBACH, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- CRONBACH, L. J. (1971). Test validation. In *measurement (2e édition)* (p. 443-507). Washington DC : American council on education.
- CRONBACH, L. J. & MEEHL, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.
- FERGUSON, G. A. (1949). On the theory of test discrimination. *Psychometrika*, 14, 61-68.
- FISCHER, C. T. (2006). *Qualitative research methods for psychologists : introduction through empirical studies*. Boston (MA) : Elsevier Academic Press.
- GUILFORD, J. P. (1946). New standards of test evaluation. *Education and Psychological Measurement*, 6, 427-438.
- GUTTMAN, L. (1954). An outline of some new methodology for social research. *The Public Opinion Quarterly*, 18, 395-404.
- GUTTMAN, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika*, 33, 469-506.
- GUTTMAN, L. (1982). Facet theory, smallest space analysis, and factor analysis. *Perceptual and Motor Skills*, 54, 487-493.
- GUTTMAN, R. & GREENBAUM, C. W. (1998). Facet theory : Its development and current status. *European Psychologist*, 3, 13-36.
- HATHAWAY, S. R. & MCKINLEY, J. C. (1943). *The Minnesota Multiphasic Personality Inventory Manual*. Minneapolis, MN : University of Minnesota Press.
- HATTIE, J. (1985). Methodological review : Assessing unidimensionality of tests and items. *Applied Psychological Measurement*, 9, 139-164.
- HATTIE, J. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement*, 20, 1-14.
- KANE, M. T. (2006). Validation. In R. L. BRENNAN (Éd.), *Educational measurement (4e édition)* (p. 17-64). Washington : American council on education-Praeger.
- KANE, M. T. (2013a). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50, 1-73. doi :[10.1111/jedm.12000](https://doi.org/10.1111/jedm.12000)
- KANE, M. T. (2013b). Validation as a pragmatic, scientific activity. *Journal of Educational Measurement*, 50, 111-122. doi :[10.1111/jedm.12007](https://doi.org/10.1111/jedm.12007)
- KELLEY, T. L. (1927). *Interpretation of educational measurements*. Yonkers-on-Hudson (NY) : World Book.
- KUHN, M. & JOHNSON, K. (2013). *Applied predictive modeling*. New York : Springer.
- LAURENCELLE, L. (1993). Et la corrélation, comment la voyez-vous? *Mesure et Évaluation en Éducation*, 16, 133-145.
- LAURENCELLE, L. (1994). Corrélation et décalage de rangs. In L. LAURENCELLE (Éd.), *Trois essais de méthodologie quantitative* (p. 1-34). Québec : Presses de l'Université du Québec.
- LAURENCELLE, L. (1997). La capacité discriminante d'un instrument de mesure. *Mesure et Évaluation en Éducation*, 20, 25-39.
- LAURENCELLE, L. (1998). *Théorie et techniques de la mesure instrumentale*. Québec : Presses de l'Université du Québec.
- LAURENCELLE, L. (2014). The discriminating capacity of a measuring instrument : Revisiting Bloom (1942)'s theory and formula. *The Quantitative Methods for Psychology*, 10, 5-12. doi :[10.20982/tqmp.10.1.p005](https://doi.org/10.20982/tqmp.10.1.p005)
- LAURENCELLE, L. (2015). Une théorie des seuils psychométriques à double contrôle d'erreur – Partie I : l'imprécision échantillonnale des centiles. *Mesure et Évaluation en Éducation*, 38, 87-109.
- LAURENCELLE, L. (2016). *L'étalonnage et la décision psychométrique (2e édition)*. Presses de l'Université du Québec : Québec.



- LAURENCELLE, L. (2017). Une théorie des seuils psychométriques à double contrôle d'erreur – Partie III : les applications illustrées. *Mesure et Évaluation en Éducation*, 40, 105-133. doi :[10.7202/1048912](https://doi.org/10.7202/1048912)
- LAURENCELLE, L. (2021). L'alpha de Cronbach, ses émules, la consistance interne, la fidélité : une mise au point. *The Quantitative Methods for Psychology*, 17, 46-50. doi :[10.20982/tqmp.17.1.p046](https://doi.org/10.20982/tqmp.17.1.p046)
- LOYE, N. (2018). Et si la validation était plus qu'une suite de procédures techniques? *Mesure et Évaluation en Éducation*, 41, 97-123. doi :[10.7202/1055898](https://doi.org/10.7202/1055898)
- MCDONALD, R. P. (1985). *Factor analysis and related methods*. Hillsdale, NJ : Erlbaum.
- MESSICK, S. (1989). Validity. In R. LINN (Éd.), *Educational Measurement* (p. 13-103). Washington, DC : American Council on Education et Macmillan.
- MISLEVY, R. J., ALMOND, R. G. & LUKAS, J. F. (2003). *A brief introduction to evidence-centered design (Research Report 03-16)*. Princeton, NJ : Educational Testing Service.
- PARISIEN, M. (2021). Validité empirique versus validité factorielle dans les inventaires de personnalité : le MMPI-2 et les échelles restructurées RC. *The Quantitative Methods for Psychology*, 17, 87-104. doi :[10.20982/tqmp.17.2.p087](https://doi.org/10.20982/tqmp.17.2.p087)
- PROULX, J. (2019). Recherches qualitatives et validités scientifiques. *Recherches qualitatives*, 38, 53-70. doi :[10.7202/1059647ar](https://doi.org/10.7202/1059647ar)
- SCHLESINGER, I. M. & GUTTMAN, L. (1969). Smallest space analysis of intelligence and achievement tests. *Psychological Bulletin*, 71, 95-100.
- SIJTSM, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika*, 74, 107-120. doi :[10.1007/s11336-008-9101-0](https://doi.org/10.1007/s11336-008-9101-0)
- SLOCUM, S. L. & ZUMBO, B. D. (2011). Assessing the unidimensionality of psychological scales : Using multiple criteria from factor analysis. *Social Indicators Research*, 102, 443-461. doi :[10.1007/s11205-010-9682-8](https://doi.org/10.1007/s11205-010-9682-8)
- STEYERBERG, E. W. (2010). *Clinical prediction models a practical approach to development*. doi :[10.1007/978-0-387-77244-8](https://doi.org/10.1007/978-0-387-77244-8)
- STOUT, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika*, 52, 589-617.
- STOUT, W. F. (1990). A new item response theory modeling approach and applications to unidimensionality assessment and ability estimation. *Psychometrika*, 55, 293-325.
- STREINER, D. L. (2003). Starting at the beginning : An Introduction to coefficient Alpha and internal consistency. *Journal of Personality Assessment*, 80, 99-103. doi :[10.1207/S15327752JPA8001_18](https://doi.org/10.1207/S15327752JPA8001_18)
- STRONG, E. K. (1927). Vocational interest test. *Educational Record*, 8, 107-121.
- TRAFIMOW, D. & EARP, B. D. (2016). Badly specified theories are not responsible for the replication crisis in social psychology : Comment on Klein. *Theory and Psychology*, 26, 540-548. doi :[10.1177/0959354316637136](https://doi.org/10.1177/0959354316637136)
- YARDLEY, L. (2008). Demonstrating validity in qualitative psychology. In J. A. SMITH (Éd.), *Qualitative psychology : A practical guide to research methods (2e édition)* (p. 235-251). Los Angeles : Sage.
- ZIEKY, M. J. (2014). An introduction to the use of evidence-centered design in test development. *Psicologia Educativa*, 20, 79-87. doi :[10.1016/j.pse.2014.11.003](https://doi.org/10.1016/j.pse.2014.11.003)
- ZINBARG, R. E., REVELLE, W., TOVEL, I. & LI, W. (2005). Cronbach's α , Revelle's β , and McDonald's ω : their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, 70, 123-133. doi :[10.1007/s11336-003-0974-7](https://doi.org/10.1007/s11336-003-0974-7)

Appendice : Tableaux sommaires des analyses factorielles

Les tableaux numérotés de 2 à 8 présentent les matrices de corrélations et les matrices de facteurs (AFP) basés chacun sur $N = 5000$ variables facettaires, celle-ci étant construites à partir des variables a , b , c et d iid de loi normale standard. **Note** : Le lecteur intéressé peut calculer lui-même la valeur asymptotique (exacte) de chaque corrélation présentée : rappelons que le coefficient r ou ρ de Pearson s'obtient en divisant la covariance des variables X et Y par la moyenne géométrique de leurs variances. Dans le cas présent, chaque variable est formée d'un ou de quelques ingrédients agrégés (par addition), eux-mêmes étant totalement indépendants. Le poids relatif de l'ingrédient dans la variable est indiqué par un chiffre en exposant : par exemple, la variable définie symboliquement par " a^3b " contient deux ingrédients, le a de poids 3 et le b de poids 1 (le poids 1 n'est pas noté). La formule symbolique à appliquer, $\rho_{X,Y} = p_{X,Y} / \sqrt{p_X \cdot p_Y}$, utilise au numérateur la somme des produits des poids (p) d'ingrédients partagés par X et Y , divisée par la racine carrée du produit des poids totaux des ingrédients des deux variables concernées. Par exemple, $\rho(ab, ac) = \rho(ab, ac)$ se développe comme $p(1a \times 1a) / \sqrt{p(1a + 1b) \cdot p(1a + 1c)} = 1 / \sqrt{2 \times 2} = 0,5$; $\rho(a^2b, ab^5c^2) = p(2a \times 1a + 1b \times 5b) / \sqrt{(2^2 + 1^2)(1^2 + 5^2 + 2^2)} = (2 \times 5) / \sqrt{5 \times 30} \approx 0,816$.

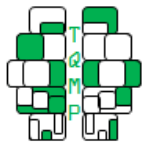


Tableau 2 ■ Matrice de corrélations et matrice des facteurs pour la structure en anneau simple (a, b, c, d)

	Matrice de corrélation				Matrice des facteurs		
	<i>a</i>	<i>b</i>	<i>c</i>	<i>d</i>	F1	F2	F3
<i>a</i>		,000	,013	−,029	,148	−,073	−,006
<i>b</i>	,000		,001	−,003	,010	,007	,034
<i>c</i>	,013	,001		−,025	,129	,084	−,008
<i>d</i>	−,029	−,003	−,025		−,190	,000	−,009

Tableau 3 ■ Matrice de corrélations et matrice des facteurs pour la structure en anneau lié (ab, ac, ad, bc, bd, cd)

	Matrice de corrélation						Matrice des facteurs				
	<i>ab</i>	<i>ac</i>	<i>ad</i>	<i>bc</i>	<i>bd</i>	<i>cd</i>	F1	F2	F3	F4	F5
<i>ab</i>		,502	,517	,490	,509	,007	,714	−,439	,231	−,494	,000
<i>ac</i>	,502		,507	,497	,003	,508	,712	,016	−,635	−,299	,000
<i>ad</i>	,517	,507		−,007	,492	,484	,705	−,564	−,201	,381	,000
<i>bc</i>	,490	,497	−,007		,501	,510	,704	,564	,201	−,381	,000
<i>bd</i>	,509	,003	,492	,501		,491	,705	−,016	,642	,302	,000
<i>cd</i>	,007	,508	,484	,510	,491		,705	,445	−,234	,500	,000

Tableau 4 ■ Matrice de corrélations et matrice des facteurs pour la structure en système simple (A, ab, ac, ad)

	Matrice de corrélation				Matrice des facteurs		
	<i>A</i>	<i>ab</i>	<i>ac</i>	<i>ad</i>	F1	F2	F3
<i>A</i>		,717	,711	,703	,981	,004	,003
<i>ab</i>	,717		,516	,509	,731	−,050	−,156
<i>ac</i>	,711	,516		,503	,724	−,117	,118
<i>ad</i>	,703	,509	,503		,716	,165	,036

Tableau 5 ■ Matrice de corrélations et matrice des facteurs pour la structure en système lié (A, ab, ac, ad, bc, bd, cd)

	Matrice de corrélation							Matrice des facteurs					
	<i>A</i>	<i>ab</i>	<i>ac</i>	<i>ad</i>	<i>bc</i>	<i>bd</i>	<i>cd</i>	F1	F2	F3	F4	F5	F6
<i>A</i>		,711	,707	,716	,007	,019	,014	,708	−,706	,010	−,008	,000	,000
<i>ab</i>	,711		,506	,515	,503	,515	,020	,794	−,210	−,333	−,463	,000	,000
<i>ac</i>	,707	,506		,507	,507	,018	,512	,790	−,201	,576	−,062	,000	,000
<i>ad</i>	,716	,515	,507		,012	,509	,502	,793	−,228	−,231	,516	,000	,000
<i>bc</i>	,007	,503	,507	,012		,506	,511	,584	,585	,228	−,514	,000	,000
<i>bd</i>	,019	,515	,018	,509	,506		,504	,592	,557	−,579	,065	,000	,000
<i>cd</i>	,014	,020	,512	,502	,511	,504		,589	,569	,332	,468	,000	,000

Tableau 6 ■ Matrice de corrélations et matrice des facteurs pour la structure en système cumulé (A, abc, abd, acd)

	Matrice de corrélation				Matrice des facteurs		
	<i>A</i>	<i>abc</i>	<i>abd</i>	<i>acd</i>	F1	F2	F3
<i>A</i>		,585	,577	,581	,724	,180	,065
<i>abc</i>	,585		,666	,670	,816	−,005	−,078
<i>abd</i>	,577	,666			,824	−,143	,091
<i>acd</i>	,581	,670	,665	1,000	,813	−,011	−,073

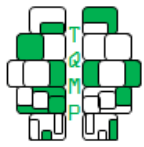


Tableau 7 ■ Matrice de corrélations et matrice des facteurs pour la structure en famille simple (A^3bcd , b c d)

	Matrice de corrélation				Matrice des facteurs		
	A^3bcd	b	c	d	F1	F2	F3
A^3bcd		,307	,305	,284	,849	,007	,012
b	,307		-,016	,013	,365	,181	,429
c	,305	-,016		-,004	,361	-,488	-,070
d	,284	,013	-,004		,326	,297	-,371

Tableau 8 ■ Matrice de corrélations et matrice des facteurs pour la structure en famille liée (A^3bcd bc , bd , cd)

	Matrice de corrélation				Matrice des facteurs		
	A^3bcd	bc	bd	cd	F1	F2	F3
A^3bcd		,410	,431	,427	,613	,210	,064
abc	,410		,504	,491	,705	-,141	,115
abd	,431	,504		,517	,714	-,027	-,026
acd	,427	,491	,517		,717	-,015	-,142

Citation

ANDRÉ, N. & LAURENCELLE, L. (2022). La validité psychométrique est d'abord un jugement : une perspective pragmatique. *The Quantitative Methods for Psychology*, 18(3), 224-240. doi : [10.20982/tqmp.18.3.p224](https://doi.org/10.20982/tqmp.18.3.p224)

Copyright © 2022, ANDRÉ et LAURENCELLE. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Received: 29/06/2022 ~ Accepted: 29/08/2022