VALIDITE DES MESURES

Marc Demeuse et Georges Henry

1. Introduction

Le terme de validité se réfère au degré selon lequel des scores de tests ou d'autres mesures prédisent ou « rendent compte » d'un ou plusieurs critères externes. On peut attribuer plusieurs significations au terme « validité », mais on doit soigneusement veiller à distinguer ce concept de celui de fidélité, sans pour autant négliger les rapports existant entre ces deux concepts. Dans ce chapitre, nous examinerons d'abord la validité en tant que telle, puis nous nous intéresserons aux rapports existants entre les deux concepts. Comme dans le cas de la fidélité, il existe plusieurs manières de considérer et définir le concept de validité.

2. Signification du terme validité

Un score est valide s'il prédit « quelque chose » et si ce « quelque chose » n'inclut pas le score lui-même. En effet, une auto-prédiction concerne la fidélité[1] et non la validité. Si deux tests sont corrélés, il existe une variance de facteur(s) commun(s). Cette variance commune augmentée de la variance spécifique est égale à la variance vraie qui est la source de la fidélité. La corrélation d'un test avec chacun des facteurs communs est son coefficient de validité par rapport à chacun de ces facteurs. Ainsi, par exemple, un test peut avoir une validité de 0,50 pour mesurer un facteur tel que N (facteur numérique des PMA de Thurstone et Thurstone, 1953) et une validité de 0,60 pour mesurer R (facteur de raisonnement des PMA).

En dehors de cette approche factorielle, on distingue un certain nombre d'acceptions au terme « validité ».

2.1. Validité prédictive[2] ou critérielle[3]

La ***validité prédictive*** repose sur la possibilité de prédire les résultats qu'obtiendront les sujets à d'autres tests ou à un instrument quelconque ou même lors d'un événement particulier, comme la réussite en fin de scolarité, la prise d'un emploi dans un délai convenu après une formation, la valeur du salaire 5 ans après la fin des études, la satisfaction dans l'emploi ou le plaisir à partager sa vie avec son conjoint, mesuré sur une échelle d'évaluation ou à travers la fréquence plus ou moins élevée de disputes...

On considère ainsi que ces résultats, externes au test à valider, constituent un critère. Cela suppose une bonne définition opérationnelle du critère et la possibilité de calculer une corrélation ou au moins une concordance entre les résultats obtenus au test à valider et la valeur du critère. C'est la valeur de cette corrélation qui indique la validité de l'instrument dans le cadre d'une prédiction.

[1] Nous avons ainsi noté la mesure de la fidélité par ρ_{tt}.
[2] En anglais, « ***predictive validity*** ».
[3] En anglais, « ***criterion validity*** ».

2.2. Validité de contenu[4]

La ***validité de contenu***, comme son nom l'indique, repose sur la nature du contenu du test par rapport à l'objet à mesurer. Dans ce type de validité, le contenu du test doit donc être en rapport direct avec ce que le test est supposé mesurer.

Ainsi, en pratique, par exemple grâce à l'analyse des programmes d'enseignement, on s'efforce de prendre en compte les matières suffisamment communes à un grand nombre d'étudiants et de donner aux épreuves une validité *a priori*, en raison du contenu même des épreuves. Les épreuves de mathématique en 6e année primaire en Communauté française de Belgique doivent prendre en compte les contenus prévus par les programmes associés à cette année d'enseignement et par les socles de compétences. Dans le domaine de la sélection professionnelle, c'est l'analyse des postes de travail à pourvoir qui fournira un ensemble de contenus aux épreuves. Par contenu, il ne faut pas nécessairement entendre des « matières » au sens scolaire classique, mais aussi des ensembles d'habiletés ou de compétences.

Certains, comme Laurencelle (1998, p. 105) qualifient ce type de « ***validité manifeste*** » : *Par validité manifeste, on désigne l'adéquation entre les contenus apparents du test et la qualité ou caractéristique qu'on souhaite mesurer (les Américains la désignent par l'expression « face validity »[5]).* Dans certaines circonstances, il est très important que le test présente une bonne validité apparente de manière à être pris au sérieux par tous ses usagers et qu'il puisse être jugé acceptable éthiquement et socialement, même si d'autres tests pourraient produire des résultats au moins aussi satisfaisants. Il en est notamment ainsi lorsque les résultats d'un test auront des répercussions importantes sur les destinées des sujets (examens de sélection, concours pour un emploi…).

2.3. Validité de construct[6] ou validité conceptuelle

La ***validité de construct*** repose sur une définition de l'objet d'évaluation lorsque celui-ci n'est pas matérialisable en termes de contenus énonçables, comme dans le cas de la validité de contenu. Les domaines de la mesure de l'intelligence ou de la personnalité constituent très certainement des exemples emblématiques : c'est la définition ou, plus exactement, la théorie relative à l'intelligence ou à la personnalité développée par des auteurs comme Spearman, Thurstone, Cattell ou Guilford qui conduit à définir opérationnellement un ensemble d'épreuves particulières, susceptibles de mesurer le niveau plus ou moins élevé atteint par un sujet particulier sur l'échelle d'intelligence ou dans certains domaines quantifiables de la personnalité. Dans certains cas, c'est même plutôt les résultats à un ensemble d'épreuves, sélectionnées sur des bases empiriques, et soumis à l'analyse factorielle qui ont conduit à la formalisation de théories.

L'analyse factorielle - il serait plus correct de dire, les différentes formes de l'analyse factorielle – a donc particulièrement contribué à investiguer les constructs liés, notamment, à la définition de l'intelligence ou aux théories de la personnalité et a conduit à formuler des modèles théoriques et opérationnels parfois très différents les uns des autres (par exemple, théories mono- ou multi-factorielle de l'intelligence, facteur général ou facteurs spécifiques…), mais sans véritables bases théoriques ou hypothèses préalables.

---

4 En anglais, « ***content validity*** ».

5 Cette expression est aussi traduite par le terme « ***validité apparente*** » (par exemple, De Landsheere, 1979, p. 290).

6 En anglais, « ***construct validity*** ».

2.4. Validité concourante[7] (ou corrélationnelle)

La ***validité concourante*** repose, comme la validité prédictive, sur la comparaison des résultats obtenus au test à valider à un critère. Dans ce cas particulier, le critère est un autre test et le principe est donc de valider un nouveau test à partir d'un autre, déjà utilisé et validé. Ce type d'approche est notamment utilisé lorsqu'il s'agit de fabriquer des formes parallèles. On peut trouver la mise en œuvre de ce type de pratique lorsqu'il convient de disposer de plusieurs formes équivalentes d'une même épreuve, par exemple dans le cas d'examens ou de concours annuels ou lorsqu'on souhaite disposer de plusieurs instruments de manière à mettre en œuvre des mesures à l'entrée à la sortie (schéma pré-test – traitement – post-test où il est nécessaire de disposer de deux instruments rigoureusement équivalents, sans risquer l'emploi d'un même instrument).

2.5. Validité incrémentale

Certains auteurs, comme Bernier et Pietrulewicz (1997, p. 219) font également référence à une forme plus rarement évoquée de validité : la ***validité incrémentale***. Celle-ci, introduite par Sechrest en 1963, trouve son intérêt lors de l'usage de batteries de tests. Dans ce cas particulier, où plusieurs tests sont administrés à un même sujet, on considérera qu'un test est valide lorsqu'il augmente de manière significative la puissance de prédiction ou d'explication de l'ensemble des tests déjà administrés. De manière opérationnelle, c'est donc la contribution partielle du test à la prédiction multiple d'un critère, à travers une régression multiple, qui détermine la validité du test. De manière plus simple, on dira dans ce cas qu'un test est valide s'il permet de mieux prédire un critère que ne le pourrait le même ensemble de tests dont il serait exclu.

3. Les procédures de validation

Bon nombre d'études expérimentales présentent des insuffisances lorsqu'on les examine du point de vue de la validité des instruments, alors que, sur le plan de la fidélité, elles sont généralement plus satisfaisantes. Il est en effet plus aisé, d'un point de vue technique, d'apporter des éléments probants quant à la fidélité, alors que la validation suppose le recours à un référent externe (définition théorique opérationnalisable, liste de contenus, autres tests déjà validés… selon la méthode de validation adoptée).

Dans les pages qui suivent, nous reprenons les considérations de De Landsheere (1982, pp. 130-134) pour décrire ces procédures de validation.

3.1. Validité prédictive ou validité critérielle

La validation est ici purement empirique : on constate si le pronostic formulé se vérifie, ou non. Il s'agit donc d'une démarche que l'on peut qualifier de préscientifique, car elle ne s'attache pas à la compréhension de la nature des phénomènes. Seul le résultat compte. Bien des tests psychotechniques utilisés jusqu'à ce jour ont été validés ainsi, de manière a-théorique.

Comment procède-t-on concrètement pour construire une épreuve de bonne validité prédictive ? Dans le domaine des sciences humaines, il est généralement beaucoup plus difficile de trouver des *critères* clairs permettant d'affirmer que telle chose s'est accomplie (par exemple, que tel enfant est devenu un bon lecteur), que de découvrir des *prédicteurs*. Nous allons donc à la fois devoir préciser la définition des prédicteurs, comme celle des

[7] En anglais, « ***concurrent validity*** » ou « ***convergent validity*** ».

critères. De plus, la mesure critérielle doit posséder quatre qualités : être en relation avec la chose prédite, ne pas être biaisée, être fidèle et pouvoir s'obtenir facilement.

a) Les critères

Dans le monde physique, on dispose souvent de critères sans ambiguïté. Si l'on prédit qu'un enfant deviendra un adulte de très grande taille, on peut décider que le critère sera : atteindre une taille supérieure à 95 % de la population; si l'on dispose de bonnes statistiques, la vérification est aisée, même si elle peut nécessiter un temps important entre la mise au point d'un test et sa validation. Si l'on prédit qu'un jeune élève fera de bonnes études universitaires, le critère est déjà beaucoup plus difficile à fixer. Ne jamais échouer dans les études supérieures, obtenir au moins une note finale donnée constituent des critères clairs. Mais sont-ils convaincants ? Bien réussir des études supérieures, n'est-ce pas plutôt acquérir un esprit scientifique ou critique, devenir un bon ingénieur, un bon professeur de langues étrangères plusieurs années après la fin des études ? ... Reste à s'entendre sur ce qu'est un « bon » professeur de langues étrangères et à s'accorder sur des comportements opérationnellement définis, qui témoigneront de l'existence des qualités choisies comme critères.

b) Les prédicteurs

La démarche générale pour choisir les prédicteurs est la suivante. On opère un ensemble de mesures, soit de comportements ou de circonstances, dont l'influence sur le phénomène à prédire ou la co-occurrence avec celui-ci a déjà été démontrée, soit encore de comportements ou de circonstances dont on peut penser qu'ils entretiennent une relation avec le phénomène. Des procédures statistiques permettent de déterminer la meilleure pondération à donner à plusieurs prédicteurs pour obtenir, de façon aussi économique que possible, une corrélation élevée avec le critère.

Répétons-le, les variables finalement sélectionnées pour la prédiction ne sont pas toujours des causes directes du phénomène prédit et, même si elles le sont, on est loin de toujours savoir comment la variable agit pour produire un phénomène donné.

Il se peut donc que des tests pronostiques contiennent des exercices qui ne semblent guère avoir de rapport avec l'objectif poursuivi; parfois aussi, les capacités testées sont tellement générales (connaissance de l'arithmétique) qu'elles peuvent prendre une forme qui, en apparence, n'a rien à voir avec le phénomène à prédire. Un test est cependant mieux accepté si l'on propose des items qui semblent directement concerner l'objet de la prédiction (*validité apparente - face validity*). Thorndike et Hagen observent, par exemple, qu'un groupe de candidats aviateurs sera plus disposé à accepter un test d'arithmétique dont les problèmes portent sur la vitesse du vent ou sur la consommation de carburant qu'un test faisant porter les mêmes types de problèmes sur l'agriculture, bien que les compétences arithmétiques évaluées soient rigoureusement identiques.

c) *Exemple*

Leclercq-Boxus (1973) a utilisé la méthode de régression multiple afin de prédire, au départ de tests administrés en dernière année de l'enseignement maternelle, les résultats en lecture lors de l'année suivante (1ère année primaire). Pour chaque enfant, les résultats à un test de lecture ont été établis après 3, 6, 9 et 12 mois de scolarité.

On a pu, ainsi établir qu'un nombre réduit de variables permet de prédire le résultat au test de lecture avec une corrélation multiple voisine (ou supérieure) à 0,80. Ce coefficient de corrélation mesure l'importance de la validité prédictive.

3.2. Validité de contenu

Un test de connaissances qui ambitionne de faire l'inventaire des acquisitions en fin d'études primaires, dans le cadre d'un programme déterminé, doit réellement couvrir les aspects importants de ce programme. Remarquons que l'appréciation de l'importance repose, soit sur un jugement de valeur, soit sur un raisonnement : les objectifs que s'est fixés l'auteur sont-ils atteints ou dans quelle mesure tel apprentissage est-il nécessaire pour accéder à tel autre, jugé important ? C'est pourquoi on parle parfois, dans ce dernier cas, de *validité rationnelle ou logique*.

Par exemple, selon que l'on considère la géométrie comme un instrument de gymnastique intellectuelle ou comme un outil destiné à résoudre des problèmes pratiques, on construira des tests de géométrie différents. Aussi, l'utilisateur devra-t-il, non seulement avoir une vision claire de ses propres conceptions, mais aussi de celles qui ont présidé à l'élaboration de l'instrument qu'il s'apprête à employer.

En ce qui concerne les comportements, les constructeurs de tests trouvent un guide précieux dans des taxonomies d'objectifs comme celles de Bloom et de ses collègues (Bloom *et al.*, 1969; Krathwohl *et al.*, 1970 ; Harrow, 1977). Même si un test porte sur tous les points importants d'une matière, on peut, en effet, considérer qu'il manque de validité de contenu s'il n'explore pas un éventail suffisant de comportements (par exemple, s'il ne fait appel qu'à la mémoire)[8].

En pratique, pour assurer la validité de contenu, on analyse les programmes, les principaux manuels utilisés, des notes de cours récentes, et l'on recueille l'avis d'enseignants, d'inspecteurs et de professeurs d'université, lorsqu'il s'agit d'épreuves scolaires. Dans le cas de tests de sélection et d'embauche, c'est l'analyse de poste et de tâches qui permet en général de définir les compétences à évaluer.

Comme la construction d'un test constitue une lourde entreprise, on s'efforce, en général, de retenir les contenus d'enseignement abordés par le plus grand nombre d'écoles. Dans les pays suffisamment outillés pour produire régulièrement de nouveaux tests nationaux ou régionaux, il arrive, cependant, que l'on introduise dans les tests des matières nouvelles ou insuffisamment étudiées, dont on souhaite que les enseignants s'y attachent spécialement. On sait, en effet, que les maîtres sont très attentifs aux questions susceptibles de déconsidérer leurs élèves ou de les déconsidérer eux-mêmes; généralement, ils insistent, par la suite, sur le domaine dans lequel ils ont été pris au dépourvu. On spécule ainsi sur l'***effet de reflux***. Le bachotage est une forme ultime d'effet de reflux: c'est le contenu du test qui finit par définir ce qui doit être appris et non plus l'inverse !

3.3. Validité de construct

Le pédagogue, comme le psychologue, explique ou décrit des comportements à l'aide d'entités ou de modèles théoriques ou hypothétiques (*constructs*) : intelligence, créativité, honnêteté... Ces entités ne sont connues qu'à travers leurs manifestations. Aussi, pour valider un test portant sur des concepts ainsi opérationnalisés, on contrôle dans quelle mesure l'épreuve recouvre les comportements qui leur sont attribués. Pour construire un test de créativité, on peut, par exemple, commencer par décrire des personnalités particulièrement créatrices ou créatives (architectes, inventeurs, artistes, etc.) et comparer leurs comportements à ceux de personnes supposées de faible créativité. Les différences observées sont, hypothétiquement,

[8] Laurencelle (1998, p. 105), à ce propos, utilise assez justement le terme « ***validité échantillonnale*** » pour indiquer que les contenus d'un test devraient constituer « un échantillon représentatif des éléments, habiletés, connaissances dont on veut mesurer la possession par un sujet ».

considérées comme les signes de la créativité. Pour valider le test, on examine s'il rend compte des caractéristiques ainsi définies.

Ce sont surtout les recherches corrélationnelles (co-occurrences de comportements) qui permettent de mettre au point de tels instruments. Par exemple, la comparaison entre sujets très créatifs et peu créatifs à laquelle on vient de faire allusion peut indiquer, d'une part, que la créativité est spécifique, c'est-à-dire qu'elle se manifeste dans un seul champ d'activité (symbolique, verbal, concret, social), et, d'autre part, qu'elle s'accompagne toujours des traits suivants assez accusés : grand pouvoir de concentration, richesse des productions divergentes dans le domaine où la créativité se manifeste : égocentrisme, rejet de la routine.

Supposons que l'on veuille construire un test de créativité verbale. Des items mettant cette aptitude en jeu sont rédigés et le test est monté. A un groupe suffisant de sujets, disons 100 élèves de 12 ans, on administre ce test et aussi des épreuves d'attention, de divergence, ainsi qu'un questionnaire de personnalité; les comportements de rejet de la routine sont observés directement et évalués à l'aide d'une échelle. On formule l'hypothèse qu'il existera : (a) une corrélation positive significative entre les scores de créativité et les scores d'attention; (b) une corrélation du même type avec la divergence et avec le caractère égocentrique; (c) une corrélation négative avec l'acceptation de la routine. C'est la vérification des hypothèses construites au départ de la définition du concept qui permet de valider l'instrument.

Il n'est pas rare qu'à titre de contrôle, on vérifie s'il y a bien absence de corrélation significative avec telle propriété apparue, mais jugée sans rapport avec la créativité par exemple, ici, l'aptitude à se servir du dictionnaire ou à effectuer rapidement des calculs mentaux[9].

Si toutes ces hypothèses se vérifient, il se confirme donc - jusqu'à preuve du contraire - que le nouveau test rend bien compte de traits caractéristiques de la créativité verbale; le *construct* est ainsi validé.

Dans la pratique, on est souvent amené à étudier les corrélations entre de nombreuses mesures. C'est pourquoi l'*analyse factorielle* est un des outils privilégiés de la validation de *construct*. Dans ce cas, les constructs sont parfois définis *a posteriori*, à partir de l'interprétation des facteurs par rapport aux items qui les constituent. Cette façon de procéder est peu satisfaisante puisqu'en quelque sorte, elle ne repose pas sur un objet défini, mais sur un ensemble éventuellement hétéroclite d'épreuves qui produisent des résultats convergents.

La démarche décrite ne donne évidemment pas de garanties absolues, puisque d'une part, il s'agit généralement d'études corrélationnelles et non causales et que, d'autre part, il n'est pas rare que les chercheurs puissent avoir omis d'investiguer certaines variables susceptibles de se manifester à travers plusieurs autres, par ailleurs étudiées.

Dans l'exemple choisi, à savoir la mesure de la créativité, il sera toujours possible d'engager, en plus, une étude longitudinale qui permettra de vérifier si, après 10, 15, 20 ans, les sujets classés parmi les personnes très créatives, le restent et occupent des fonctions ou pratiques des hobbies réputés créatifs. Dans ce cas, la validité prédictive vient s'ajouter à la validité du *construct*.

[9] L'absence de corrélation significative ne signifie pas que les sujets créatifs sont, moins que les autres, capables de se servir d'un dictionnaire ou d'effectuer des calculs mentaux, mais que ces deux aptitudes ne sont pas d'autant plus développées que la mesure de la créativité est élevée. C'est une corrélation négative entre le score de créativité et d'autres mesures qui conduiraient à affirmer que créativité et calcul mental ou usage du dictionnaire sont en quelque sorte en concurrence chez les sujets.

3.4. Validité concourante

L'effort théorique inhérent à la véritable validation de *construct* d'un test est souvent considérable. Parfois, la validité de *construct* d'un test semble si fermement établie que ce test devient une sorte d'intermédiaire de validation. Par exemple, après plusieurs décennies d'utilisation, certains continuent à penser que le test d'intelligence générale de Raven reste valide. On peut imaginer qu'un chercheur, estimant le test de Raven trop lourd à administrer, tente de mettre au point une épreuve beaucoup plus économique. La validité pourrait alors être évaluée en calculant la corrélation entre les résultats obtenus aux deux tests, par les mêmes sujets (validité concourante). Le danger de pareille façon de procéder est évident et beaucoup de constructeurs n'y ont pas échappé. Choisissant la solution de facilité que ce type de validation corrélationnelle constitue, ils prolongent simplement des démarches erronées.

4. Problèmes spécifiques

4.1. Problèmes liés à la prédiction multiple

On constate souvent qu'après introduction de trois ou quatre tests dans l'équation de régression, l'addition de nouveaux tests apporte peu d'amélioration (absence de validité incrémentale) car les facteurs communs sont couverts par les tests déjà pris en compte. C'est pourquoi, lorsque des tests appartenant à une batterie présentent une inter-corrélation élevée, on les combinera souvent en un seul score avant de procéder aux analyses de régression. Sinon, on constate l'apparition de phénomènes de colinéarité[10].

On doit aussi être attentif aux phénomènes d'idiosyncrasie à l'échantillon qui entraînent une surestimation du coefficient de validité résultant de corrélations accidentelles. Ceci est particulièrement le cas lorsque les échantillons ont une taille faible et possèdent des caractéristiques très particulières[11]. Lorsque les coefficients sont recalculés, au départ d'un autre échantillon, les formules peuvent ne plus manifester un pouvoir prédictif aussi élevé. Il y a donc toujours nécessité de procéder à des études de validité croisée. Pour ce faire, on divise, aléatoirement, l'échantillon de départ en deux sous-échantillons. Sur le premier sous-échantillon, on calcule l'équation prédictive (avec le coefficient de corrélation multiple). On applique cette formule sur le deuxième sous-échantillon et on vérifie s'il n'y a pas baisse anormale de la validité prédictive. On effectue, ensuite, la même démarche en prenant le deuxième sous-échantillon comme point de départ et on établit l'ampleur de la baisse moyenne de validité.

4.2. La correction pour atténuation

Soient deux tests T1 et T2 qui mesurent exactement les mêmes contenus, les mêmes processus ou les mêmes compétences. Dans ce cas, la composante « vraie » (X_∞) est identique, mais entachée d'une erreur différente, propre à chaque test et notée, respectivement e_1 et e_2.

---

[10] Il y a colinéarité exacte lorsqu'un prédicteur est une combinaison linéaire d'un ou plusieurs autres prédicteurs. La prise en considération de ces prédicteurs liés entre eux est techniquement problématique et n'apporte aucun source supplémentaire d'information. C'est pourquoi on agrège généralement ces données ou on élimine celles qui sont liées parfaitement de manière à ne conserver qu'un seul prédicteur.

[11] Ces caractéristiques peuvent être fortuites, dans le cas d'un échantillon aléatoire et simple de très petite taille, mais sont beaucoup plus probables dans le cas d'échantillons non probabilistes. Par exemple, la mise au point d'un instrument sur un échantillon d'étudiants en psychologie alors qu'il est destiné à la population générale. Le cumul des deux biais - petite taille et constitution non probabiliste - conduit avec encore plus de certitude à des idiosyncrasies.

$$\mathbf{X_{t1} = X_{\infty} + e_1}$$

$$\mathbf{X_{t2} = X_{\infty} + e_2}$$

Les scores observés (X_{t1} et X_{t2}) sont donc différents et la corrélation entre ces deux scores observés ne sera pas égale à 1, bien que les scores vrais soient identiques, puisque les erreurs ne sont pas corrélées (ρ_{e1e2} = **0)**. C'est la variance d'erreur qui intervient dans les deux mesures qui a pour effet d'abaisser la corrélation entre les scores observés. Si on veut connaître la corrélation entre les deux composantes vraies, on doit effectuer une correction afin de neutraliser l'effet de ces variances d'erreurs.

On peut estimer la corrélation entre les scores vrais de la manière suivante :

$$\rho_{\infty\omega} = \frac{\rho_{t1t2}}{\sqrt{\rho_{t1t1}\rho_{t2t2}}}$$

où $\rho_{\infty\omega}$ est la corrélation entre les composantes vraies des deux tests (on les indicera ∞ et ω de manière à les distinguer),

ρ_{t1t2} est la corrélation entre les scores observés et

ρ_{t1t1} et ρ_{t2t2} sont les coefficients de fidélité des tests T1 et T2.

Le problème se pose, néanmoins, de savoir quel type de fidélité doit être prise en compte pour effectuer ce calcul. En fait, il convient de se placer dans les mêmes conditions que celles qui ont permis de calculer ρ_{t1t2}. On choisira donc le plus souvent, la technique des formes parallèles à peu de temps de distance, ce qui correspond bien à l'utilisation de deux tests supposés mesurer la même chose et administrés à deux moments proches. Par contre, si on utilise un coefficient de consistance interne, il y aura risque de sous-correction, ce type de fidélité fournissant souvent des coefficients de fidélité plus élevés (du fait notamment qu'il n'y a, pour chaque test, qu'une seule administration et donc pas de fluctuation des résultats dans le temps).

Par ailleurs, lorsqu'on se place dans une perspective prédictive, ce qu'on veut, c'est connaître la capacité du test de prédire le critère. La variance d'erreur existant dans le test prédictif doit être prise en compte (puisqu'elle peut varier de test à test), mais non celle existant dans le critère puisqu'on sera toujours amené à prédire un même critère manquant éventuellement de fidélité. Dans ce cas, on utilisera la formule suivante

$$\rho_{x\omega} = \frac{\rho_{xy}}{\sqrt{\rho_{yy}}}$$

où $\rho_{x\omega}$ est la corrélation corrigée pour la variance d'erreur existant dans le critère y et ρ_{xy} est la corrélation entre le score au test x et le critère y.

Un problème supplémentaire se pose néanmoins: ρ_{xy} n'est généralement pas un paramètre de population, c'est-à-dire la vraie corrélation existant au niveau de la population. C'est une statistique d'échantillon. Il existe donc des variations possibles, qui peuvent parfois être importantes et, par conséquent, on peut corriger un coefficient déjà surfait. A ce moment, $\rho_{x\omega}$

et $\rho_{\infty\omega}$ pourraient devenir supérieurs à **1**, ce qui est manifestement absurde, puisqu'il s'agit de coefficients de corrélation. Pour pouvoir utiliser ce genre de correction, il faut donc s'assurer que l'***erreur d'échantillonnage*** sur les coefficients de fidélité est faible. On exigera donc des effectifs de grande taille lorsqu'il s'agit d'échantillon, ce qui garantit de faibles erreurs standard d'échantillonnage et minimise le risque de sur-correction conduisant à des résultats absurdes.

4.3. Relations entre validité et longueur d'un test

Puisque la validité est liée à la fidélité et que la fidélité est elle-même liée à la longueur des tests, il existe donc un lien entre la validité d'un test et sa longueur. Si on multiplie par m la longueur d'un test x, sa validité par rapport à un critère s'obtient par la formule suivante :

$$\rho_{mx.y} = \frac{\rho_{xy}}{\sqrt{\frac{1-\rho_{xx}}{m} + \rho_{xx}}}$$

où $\rho_{mx.y}$ est la validité d'un test m fois plus long que le test x initial ;

ρ_{xy} est la corrélation entre le score au test x et le critère y, c'est-à-dire la validité du test initial ;

ρ_{xx} est la fidélité du test x initial ;

m est le coefficient d'allongement du test (par exemple, m = 2 si on double la longueur du test, m = 0,5 si on diminue le test de moitié).

Cette formule permet d'estimer la validité d'un test qui résulterait de l'allongement (ou du raccourcissement) d'un test dont les caractéristiques sont déjà connues. Cela permet, par exemple, d'étudier l'impact de la suppression de certains items parce que le test est trop long dans les circonstances normales d'utilisation (à condition que les items supprimés soient bien équivalents à ceux qui sont maintenus). On peut également résoudre, avec cette formule, un autre type de problèmes : ayant pré-testé un instrument, dont la validité est jugée insuffisante, quelle devrait être la longueur d'un nouveau test (combien d'items devraient être ajoutés ?) pour que le test présente la validité souhaitée ?

4.4. Effet de la dispersion des aptitudes sur la validité

Lorsque la dispersion des aptitudes est pratiquement nulle, la corrélation avec le critère est nécessairement très faible. Ce constat se déduit aisément de la formule qui permet de calculer une corrélation. Par conséquent, un test peut avoir une bonne validité pour un groupe de sujets dont les aptitudes sont très variées et avoir une faible validité pour un groupe de sujets dont les aptitudes sont très semblables.

Un problème particulier se pose, si on veut procéder à des études de validation, dans le cas de concours ou de sélection professionnelle. En effet, les expériences de validation ne peuvent se faire que sur les candidats admis, or dans ce cas, les sujets admis auront des aptitudes moins étalées que dans la population qui a présenté l'examen d'entrée ou le concours de sélection. On a en effet éliminé les sujets supposés les moins aptes, selon les résultats du test. Il est dès lors impossible de savoir si les sujets non sélectionnés auraient ou non réussi leur formation ou auraient occupé leur fonction de manière satisfaisante ou non, puisque par définition, on ne

leur a pas permis de poursuivre. La validité calculée sur cette population sélectionnée sera donc généralement plus faible que la validité réelle du test impossible à estimer.

Bibliographie

Bernier, J.J., Pietrulewicz, B. (1997). *La psychométrie. Traité de mesure appliquée.* Montréal : Gaëtan Morin éditeur.

Bloom, B.S., Engelhart, M.D, Furst, E.J., Hill, W.H., Krathwohl, D.R. (1969). *Taxonomie des objectifs pédagogiques. Domaine cognitif.* Montréal : Education nouvelle.

De Landsheere, G. (1979). *Dictionnaire de l'évaluation et de la recherche en éducation.* Paris : Presses universitaires de France.

De Landsheere, G. (1982). *Introduction à la Recherche en Education.* Liège : Editions G. Thone.

Harrow, A.J. (1977). *Taxonomie des objectifs pédagogiques. Domaine psychomoteur.* Montréal : Presses de l'Université du Québec.

Krathwohl, D.R., Bloom, B.S., Masia, B.B. (1970). *Taxonomie des objectifs pédagogiques. Domaine affectif.* Montréal : Education nouvelle.

Leclercq-Boxus, E. (1973). Etude différentielle de la prédiction du rendement en lecture en première année primaire. In *Recherches sur les handicaps socio-culturels.* Ministère de l'Education nationale, Organisation des Etudes. Bruxelles. 197-210.

Thurstone, L.L., Thurstone, T.G. (1953). *Batterie factorielle P.M.A. Adaptation française de SRA Primary Mental Abilities Intermediate 11-17. Signification Verbale, Spatial, Raisonnement, Numérique, W - Fluidité verbale.* Paris: Centre de Psychologie appliquée (pour la version française).