Un regard panoramique sur la validité de l'évaluation en sciences de la santé

A panoramic perspective on the validity of assessment in health sciences

Um olhar panorâmico sobre a validade da avaliação em ciências da saúde

Jean-Sébastien Renaud
ID ORCID : 0000-0002-2816-0773
Faculté de médecine, Université Laval, Québec, Canada
VITAM, Université Laval, Québec, Canada

Éric Dionne
ID ORCID : 0000-0003-3046-4386
Faculté d'éducation, Université d'Ottawa, Ottawa, Canada

Miriam Lacasse
ID ORCID : 0000-0002-2981-0942
Faculté de médecine, Université Laval, Québec, Canada

Julie Thériault
CIUSSS Capitale nationale, Québec, Canada

Mots clés : évaluation, pédagogie des sciences de la santé, validation, validité

La validité est la qualité la plus importante d'une évaluation. En revanche, c'est souvent un concept peu maîtrisé par le corps enseignant en sciences de la santé (professeurs, cliniciens-enseignants, etc.). Cet article vise à aider le corps enseignant en sciences de la santé à comprendre comment documenter ou juger la validité d'une évaluation. Il aborde la validation d'une évaluation *par des méthodes qualitatives et quantitatives selon l'approche la plus reconnue, soit celle décrite dans les Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014). Nous décrivons les différents types de preuves de validité, tout en expliquant comment il est possible de les documenter dans un contexte d'évaluation en sciences de la santé.*

Key words: assessment, health sciences education, validation, validity

Validity is the most important quality of an assessment. However, it is often a poorly understood concept among health sciences educators (professors, clinical educators, etc.). This article aims to help health sciences educators understand how to document or evaluate the validity of an assessment. It addresses the validation of an assessment using qualitative and quantitative methods from the perspective of the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014), the most widely recognized approach. The different sources of validity evidence are described, along with explanations of how they can be documented in the context of health sciences education.

Palavras-chave: avaliação, pedagogia das ciências da saúde, validação, validade

A validade é a qualidade mais importante de uma avaliação. No entanto, trata-se frequentemente de um conceito que o corpo docente das ciências da saúde (professores, clínicos-docentes, etc.) domina pouco. Este artigo tem como objetivo ajudar os docentes desta área a compreender como documentar ou julgar a validade de uma avaliação. Aborda a validação de uma avaliação através de métodos qualitativos e quantitativos, segundo a abordagem mais reconhecida, descrita nos Standards for Educational and Psychological Testing (AERA, APA & NCME, 2014). Descrevemos os diferentes tipos de provas de validade, explicando-se de que forma podem ser documentados no contexto da avaliação em ciências da saúde.

Note des auteurs : la correspondance liée à cet article peut être adressée à jean-sebastien.renaud@fmed.ulaval.ca

Introduction

Le corps enseignant (professeurs, cliniciens-enseignants, etc.) et les directions de programmes en sciences de la santé ont la responsabilité d'évaluer chez les étudiants les compétences visées par le programme de formation. Souvent, ils doivent aussi évaluer les candidats désireux d'être admis dans le programme. Par souci d'éthique professionnelle (Jeffrey, 2013), mais aussi pour respecter les critères d'agrément des programmes et les politiques d'évaluation des institutions universitaires (Comité d'agrément des facultés de médecine du Canada, 2023 ; Vice-rectorat aux études et aux affaires étudiantes, 2025), les enseignants et les responsables de programme ont le devoir de porter un regard sur la validité des évaluations réalisées (Marceau et al., 2022), qui est la qualité la plus importante d'une évaluation (Downing & Haladyna, 2009). En revanche, la validité est souvent un concept peu maîtrisé par les responsables de formation en sciences de la santé (p. ex., corps enseignant, directions de programmes). Qui plus est, le concept de validité a passablement évolué depuis son apparition il y a plus d'un siècle (André et al., 2015 ; Brennan, 2006 ; Kane, 2013 ; Loye, 2018). Cette évolution a généré plusieurs conceptualisations de la validité et ces dernières, mêmes celles jugées désuètes, s'observent toujours dans les discours et les pratiques. Il en résulte parfois des incompréhensions entre les différents intervenants s'intéressant à la validité de l'évaluation en sciences de la santé (experts en évaluation, membres du corps enseignant, directeurs de programmes, chercheurs, etc.) ainsi que des écarts entre les pratiques recommandées de validation et celles adoptées en pratique (St-Onge & Young, 2015 ; St-Onge et al., 2017).

Dans ce contexte, cet article vise à aider le corps enseignant, les directions de programmes et les autres intervenants en sciences de la santé à comprendre comment documenter ou juger la validité d'une évaluation. Bien qu'il existe d'autres références visant à démystifier la validité auprès de cette population (p. ex., Cook & Beckman, 2006 ; Downing, 2003), elles sont plutôt sommaires et aucune n'est en français. Reconnaissant ce besoin, cet article propose une présentation de nature pédagogique, en

français, du concept de validité et du processus de validation en contexte de formation dans les programmes en sciences de la santé. Plus précisément, il présente l'approche contemporaine la plus connue et reconnue (Lineberry, 2020), soit celle que proposent les *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), ci-après nommés les *Standards* 2014 pour alléger le texte. Cet ouvrage de référence a été élaboré conjointement par trois organismes américains, l'American Education Research Association (AERA), l'American Psychological Association (APA)et le National Council on Measurement in Education (NCME) et reflète la vision dominante sur les bonnes pratiques d'évaluation en éducation et en psychologie.

L'article débute en définissant le concept de validité et en situant son importance. Il poursuit avec son opérationnalisation, où les différents types de preuves de validité sont décrits. Enfin, il se conclut par une discussion sur la validation en pratique et sur les différentes limites et critiques liées au concept de validité.

La définition de la validité

La dernière édition des *Standards*, publiée en 2014, définit la validité de la manière suivante : c'est « le degré auquel les preuves et la théorie soutiennent l'interprétation des résultats à un test dans le contexte de l'utilisation qui en est fait » (p. 11, traduction libre). À noter que le mot « test » est utilisé dans les *Standards* 2014 dans le même sens que « évaluation » et, comme le fait Downing (2003), ces deux termes seront utilisés de manière interchangeable dans cet article. Cette définition met de l'avant le fait que la validité est une question de degré et que celui-ci dépend du nombre et du poids des preuves empiriques ainsi que de la théorie soutenant l'interprétation des résultats au test dans un contexte d'utilisation. De plus, la validité concerne les interprétations faites à partir des résultats au test et non le test lui-même. Puisqu'elle concerne le caractère adéquat de l'interprétation des scores dans un contexte d'utilisation donné, ceci implique deux choses. D'abord, l'interprétation ou les interprétations doivent être précisées clairement pour qu'il soit possible de les valider. Ensuite, la validité doit être démontrée pour chaque contexte d'utilisation et pour chaque nouvelle interprétation proposée. Par exemple, un test évaluant le raisonnement clinique en médecine ne peut être utilisé en physiothérapie sans que cette nouvelle utilisation du test ait fait l'objet d'une validation.

Toujours selon les *Standards* 2014, la validation fait référence au « processus d'accumulation de preuves de manière à avoir une base scientifique solide pour soutenir l'interprétation proposée des résultats au test » (p. 11, traduction libre). Ces preuves sont classées en cinq catégories : 1) les preuves basées sur le contenu du test, 2) les preuves basées sur le processus de réponse, 3) les preuves basées sur la structure interne du test, 4) les preuves basées sur les relations avec d'autres variables et 5) les preuves basées sur les conséquences du test. Ces preuves, qui peuvent soutenir la validité du test ou non, selon le cas, doivent être intégrées dans un tout cohérent pour former l'argumentaire de validité.

L'importance de la validité

La validité est considérée comme la qualité la plus importante d'une évaluation (Downing & Haladyna, 2009). Si l'interprétation et l'utilisation des résultats à un test ne peuvent être justifiées, ce dernier n'a aucun sens et aucune utilité. En conséquence, la validité est directement associée à la crédibilité d'une évaluation et, en retour, à son acceptabilité et à ce qui en découle (décision, rétroaction, classement, etc.).

L'utilisation d'un cadre conceptuel de la validité a également son importance. D'abord, un tel cadre permet de donner un sens au concept de validité. Ensuite, il permet de guider et de justifier la démarche de validation. Enfin, il sert à intégrer et à interpréter les résultats de cette démarche. Pour ces raisons, il est préférable de spécifier le cadre conceptuel de la validité utilisé pour guider la démarche de validation puisque différents cadres mènent à différentes démarches.

Bref, démontrer la validité d'une évaluation permet de soutenir l'interprétation des résultats au test et, conséquemment, sa crédibilité et son acceptabilité. Le cadre conceptuel de la validité permet, de son côté, d'éclairer le processus de validation.

Les preuves de validité

Selon les *Standards* 2014, cinq types de preuves peuvent être recueillies dans un processus de validation. Voici plus concrètement comment documenter la validité d'une évaluation à partir de ces cinq types de preuves.

Les preuves basées sur le contenu du test

Les preuves de validité basées sur le contenu du test sont « obtenues en analysant la relation entre le contenu du test et le construit qu'il tente de mesurer » (AERA et al., 2014, p. 14, traduction libre). Elles se rapportent

à la cohérence entre ce que les concepteurs du test cherchent à évaluer et le contenu du test. Un prérequis à l'évaluation du contenu est donc que le construit à évaluer par le test soit clairement défini. Ce construit peut aussi bien être un ensemble d'objectifs d'apprentissage qu'un concept psychologique. En éducation, le terme « domaine » est souvent utilisé pour désigner l'ensemble des apprentissages que cherche à évaluer un test.

Il n'existe pas de règle claire et préétablie pour juger la pertinence du contenu d'une évaluation. C'est une démarche qui repose sur le jugement d'une ou de plusieurs personnes (enseignants, experts de contenu, spécialistes de l'évaluation, etc.) et cette démarche a plus ou moins d'ampleur, selon l'importance de l'évaluation.

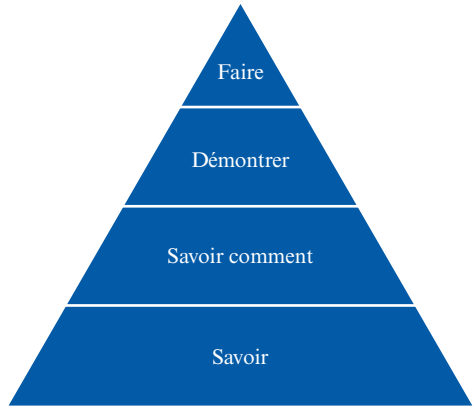
Plusieurs aspects du test peuvent être examinés pour apprécier la pertinence du contenu d'une évaluation (AERA et al., 2014 ; Downing, 2003) et ces derniers peuvent être regroupés en trois catégories : 1) la cohérence entre le construit évalué et la méthode d'évaluation utilisée, 2) la représentativité du contenu du test par rapport au construit évalué et 3) la qualité des items utilisés. Ces trois catégories sont abordées ci-après.

La cohérence entre le construit et la méthode d'évaluation

Examiner la cohérence entre le construit évalué et la méthode d'évaluation nécessite de poser un jugement sur le choix du type de test et d'items. Bref, il faut que la méthode d'évaluation convienne à ce que l'on souhaite évaluer (Yudkowsky et al., 2020). Par exemple, pour évaluer si les étudiants maîtrisent l'examen physique d'un patient, il faudra orienter l'évaluation vers un examen pratique plutôt que vers un examen écrit. Différents cadres conceptuels, seuls ou combinés, peuvent être utilisés pour y arriver, comme celui de Bloom (1956) pour les objectifs cognitifs et celui de Krathwohl et al. (1964) pour les objectifs affectifs. En sciences de la santé, la pyramide de Miller (1990), présentée à la figure 1, est souvent retenue comme cadre conceptuel pour faire cet exercice. Dans les prochains paragraphes, les niveaux de cette pyramide seront expliqués à partir des textes de Miller (1990) et de Yudkowski et al. (2020).

Le bas de la pyramide correspond aux savoirs. Il s'agit des connaissances factuelles et conceptuelles, telles que les systèmes du corps humain, les pathologies et leurs symptômes, les divers principes biochimiques, etc., pour lesquelles les évaluations écrites conviennent bien, comme les questions à réponse choisie (questions à choix multiples, de type vrai/faux, d'association) ou à réponse construite courte.

Figure 1
Pyramide de Miller



Note. Adaptée de *The assessment of clinical skills/competence/performance* par G. E. Miller, 1990, *Academic Medicine*, *65*(9), S63-67.

Le second niveau de la pyramide de Miller se nomme le savoir comment (*Knows how*). Il porte sur la capacité d'utiliser ses connaissances et, au besoin, d'aller chercher les informations manquantes, pour résoudre des problèmes. Il s'agit donc d'un niveau cognitif supérieur, nécessitant un traitement d'information plus complexe, par exemple l'intégration de plusieurs concepts ou connaissances. Ce niveau d'apprentissage peut également être évalué par des tests écrits. Cependant, il faut faire attention à ce que les items n'évaluent pas simplement les savoirs mais bien l'utilisation de ces savoirs pour résoudre des problèmes. Les examens oraux et les essais peuvent aussi être utilisés.

Le niveau suivant est la démonstration des apprentissages (*Shows how*). Ce niveau correspond à la capacité des étudiants à mettre en pratique leurs apprentissages en situation contrôlée. Autrement dit, en contexte académique, souvent lors d'une simulation (patient simulé, examen objectif clinique structuré, etc.), l'évaluation vérifie si l'étudiant arrive à mettre en pratique ce qu'il a appris, par exemple ausculter un patient, communiquer une mauvaise nouvelle à un patient, procéder à une anamnèse, etc. L'appréciation par simulation, les examens cliniques objectifs structurés et les examens avec patients simulés sont souvent utilisés dans ce contexte.

Le niveau supérieur de la pyramide de Miller va encore plus loin et cherche à vérifier si, en situation clinique réelle, dans une pratique autonome, l'étudiant agit avec compétence. C'est ce niveau qui est généralement évalué en contexte de stages cliniques, au moyen d'outils comme les mini-CEX, les évaluations multisources, aussi appelées évaluations 360, l'observation directe des habiletés cliniques (*direct observation of clinical skills* ou *DOPS*), les portfolios, les fiches d'évaluation de fin de stage (*in-training evaluation reports* ou *ITER*), la supervision par discussion de cas, etc.

En somme, il faut évaluer la cohérence entre le construit évalué et la méthode d'évaluation choisie afin de s'assurer que la tâche demandée à l'étudiant permet d'évaluer l'objectif d'apprentissage ciblé. Par exemple, un test de connaissances, aussi bon soit-il, peut difficilement évaluer la capacité d'un étudiant à appliquer ses connaissances en situation clinique simulée ou réelle. Une note s'impose toutefois. Il faut éviter d'associer automatiquement une méthode (p. ex., les questions à choix multiples) à un niveau de la pyramide de Miller. Ce sont des principes généraux qui sont donnés ici comme exemples et il est nécessaire d'examiner la tâche demandée aux répondants pour bien déterminer le niveau d'apprentissage évalué. En effet, il arrive que des questions à choix multiples évaluent des processus cognitifs complexes et une évaluation en stage pourrait n'évaluer que des connaissances.

La représentativité du contenu du test

Il est également important d'examiner la représentativité du contenu du test par rapport au construit évalué pour juger de la pertinence du contenu. Il est reconnu depuis longtemps (p. ex., Messick, 1989) qu'une des menaces sérieuses à la validité d'une évaluation est la sous-représentation du contenu à évaluer (*construct underrespresentation*). Lorsque c'est le cas, le contenu évalué par le test est trop restreint, c'est-à-dire qu'il couvre de manière insuffisante les objectifs d'apprentissage à évaluer ou encore, chaque objectif n'est pas suffisamment évalué. Par conséquent, les résultats au test ne permettent pas de tirer des conclusions sur l'ensemble des objectifs d'apprentissage que visait le test. Une autre menace à la représentativité du contenu est la présence d'items (questions d'examen, tâches, etc.) qui génèrent de la variance non attribuable à la caractéristique que l'on tente d'évaluer (AERA et al., 2014). C'est le cas, par exemple, lorsqu'une question d'examen est plus difficile à comprendre et à réussir pour les personnes n'appartenant pas à la culture dominante.

Ainsi, en pratique, il faut répondre à au moins trois questions pour vérifier la représentativité du contenu d'une évaluation : 1) Est-ce que tous les objectifs d'apprentissage visés par le test sont effectivement évalués par celui-ci ? 2) Est-ce que le nombre d'items évaluant chaque objectif d'apprentissage est représentatif de l'importance relative de ces objectifs ? et, enfin, 3) Est-ce que les items choisis sont pertinents, c'est-à-dire, est-ce qu'ils évaluent effectivement ce qu'ils doivent évaluer ?

La table de spécification, aussi appelée tableau directeur ou *blueprint*, permet de répondre aux deux premières questions en détaillant les objectifs d'apprentissage évalués, le nombre d'items rattachés à chacun d'eux et le ou les niveaux cognitifs évalués pour chaque objectif (p. ex., en se référant aux taxonomies des objectifs d'apprentissage). La table de spécification est abordée de façon approfondie dans plusieurs autres ouvrages (Coderre et al., 2009 ; Fives & DiDonato-Barnes, 2013 ; Laurier et al., 2005).

La troisième question, quant à elle, nécessite généralement qu'un ou plusieurs experts sur le construit évalué, qualifiés d'experts de contenu, portent un jugement sur la force du lien entre les items utilisés et le construit évalué (AERA et al., 2014 ; Downing, 2003). Ceci peut se faire de plusieurs manières. Encore une fois, l'ampleur de la démarche varie selon l'importance du test. Pour une évaluation en classe à enjeux faibles, demander l'avis de collègues qualifiés pour se prononcer sur la pertinence des items utilisés peut suffire. Toutefois, lorsque les enjeux sont élevés, plusieurs experts sont appelés à se prononcer et leur avis peut être sollicité de différentes manières. Dans certains cas, ces experts sont consultés individuellement afin de recueillir leur avis sur la pertinence des items. Une méthodologie Delphi, qui est une démarche systématique de consultation d'experts visant à dégager un consensus sur un sujet donné, peut également être utilisée. Il arrive aussi que des indices statistiques, tels que les coefficients kappa et alpha de Cronbach, soient utilisés pour déterminer le degré d'accord des experts sur la pertinence des items du test (Polit & Beck, 2006). L'indice de validité de contenu des items (*content validity index* ou *CVI*) (Lynn, 1986) peut aussi servir pour déterminer la pertinence de chaque item selon un groupe d'experts. Bien que nécessaire, l'avis d'experts comporte des limites. Par exemple, il leur est généralement difficile de prédire la difficulté des questions (Kibble & Johnson, 2011), bien que nous voulions généralement éviter les questions trop difficiles ou trop faciles pour les candidats ciblés par l'évaluation.

La qualité des items

Une autre preuve de validité relative au contenu d'une évaluation est la qualité des items utilisés (questions, cas cliniques, etc.) (Downing & Haladyna, 2009). Ils doivent être clairs et rédigés en respectant les lignes directrices pertinentes, ce qui augmentera leur probabilité d'être de bons items (Abozaid et al., 2017). Par exemple, Downing (2009b, p. 158-159) résume les lignes directrices courantes pour la rédaction de questions à choix multiples et Johnson et Morgan (2016, p. 50-68), celles pour la rédaction d'items de questionnaires. La qualité des items dépend aussi de la qualification des personnes qui les ont rédigés. En effet, on s'attend à ce que les personnes responsables de rédiger les items soient des experts du contenu évalué.

Les preuves basées sur le processus de réponse

L'évaluation du processus de réponse fait référence à l'étude de la correspondance entre le construit évalué et le processus cognitif des répondants et des évaluateurs (Cook & Beckman, 2006). Il faut vérifier si les étudiants comprennent bien les items et s'ils utilisent les processus cognitifs que l'on cherche à évaluer. Par exemple, pour évaluer la capacité des étudiants à générer des hypothèses diagnostiques probables pour un ensemble de symptômes cliniques, il faut s'assurer que c'est bien la tâche cognitive que réalisent les étudiants lorsqu'ils répondent aux items. Du côté des évaluateurs, il importe de s'assurer qu'ils comprennent et appliquent les critères d'évaluation de manière appropriée.

En pratique, l'évaluation de la correspondance entre le construit évalué et le processus cognitif des répondants et celui des évaluateurs s'effectue le plus souvent au moyen de méthodes de nature qualitative. Les *Standards* 2014 donnent comme exemple la conservation et l'analyse des traces laissées par les étudiants en effectuant la tâche d'évaluation demandée, notamment dans le cas d'une évaluation écrite (brouillons, suivis de modifications dans les logiciels de traitement de texte, etc.). Ils citent aussi diverses autres stratégies, telles que le suivi des mouvements des yeux, le cas échéant, l'évaluation du temps de réponse et l'étude des relations entre les parties du test et entre le test et d'autres variables. Ils mentionnent également qu'il est possible de questionner directement les étudiants qui répondent au test pour qu'ils explicitent leur processus de réponse aux items. Cette stratégie est sans doute l'une des plus fréquemment utilisées pour documenter le processus de réponse, du moins en éducation, et pour la validation de

questionnaires. Elle est suggérée par différents auteurs comme DeVellis (2012) ou Drennan, (2003). La méthode de la pensée à voix haute (*think-aloud protocol*) peut être employée pour amener les répondants au test à expliciter leur processus de réponse aux items ou encore, pour préciser le processus qu'utilisent les évaluateurs pour noter les candidats.

Enfin, certains auteurs (Downing, 2003 ; Downing & Haladyna, 2009) ajoutent à l'évaluation du processus de réponse le contrôle de l'intégrité des données d'évaluation. Cette intégrité peut être compromise par des erreurs de diverses natures, comme des erreurs de transcription des données, des erreurs dans la clé de correction ou dans la correction, des problèmes de lecture des grilles à bulles avec un lecteur optique, etc.

Les preuves basées sur la structure interne

Selon les *Standards* 2014, « l'analyse de la structure interne d'un test peut indiquer dans quelle mesure les relations entre les items et celles entre les composantes du test sont conformes au construit sur lequel les interprétations du test sont basées », (p. 16, traduction libre). Les outils statistiques utilisés pour réaliser cette analyse peuvent varier selon la théorie de la mesure utilisée. Il en existe trois, soit la théorie classique des tests (TCT), la théorie des réponses aux items (TRI) et la théorie de la généralisabilité (GEN).

Le présent article discute des analyses fréquemment réalisées dans le contexte de la TCT étant donné que c'est la plus ancienne, la plus connue et la plus populaire des trois théories de la mesure. Comme le résume bien De Champlain (2010) dans son introduction à la TCT, cette dernière postule que le score observé à un test pour une personne (X) est composé de son score vrai (V) et de l'erreur de mesure aléatoire (E). La TCT propose donc une relation linéaire entre ces variables, où $X = V + E$. Le score vrai est défini comme l'espérance mathématique du score observée sur un nombre infini d'administrations du même test à cette personne. Quant à l'erreur de mesure aléatoire, elle peut être définie comme tout événement ou circonstance (distractions, motivation, anxiété, etc.) qui corrompt la performance optimale d'une personne évaluée (Osterlind, 2010, p. 62). L'espérance mathématique de l'erreur de mesure aléatoire est de 0. Autrement dit, l'erreur de mesure aléatoire pour une personne est parfois positive, parfois négative, selon l'administration du test, mais sur un nombre infini d'administrations d'un même test, ces erreurs viennent à s'annuler. Dans la situation hypothétique d'un test sans erreur aléatoire de mesure, le score vrai serait égal au score observé.

Les paragraphes suivants présentent les analyses typiques dans le contexte de la TCT, soit :

1) L'analyse de la fidélité et de la précision des scores (p. ex., Dore et al., 2010)
2) L'analyse de la structure factorielle (p. ex., Hoban et al., 2005)
3) L'analyse d'items (p. ex., Yang et al., 2011)
4) L'analyse du fonctionnement différentiel d'items (p. ex., Roberts et al., 2009)

L'analyse de la fidélité

L'analyse de la fidélité permet de juger de la reproductibilité des scores. La fidélité des scores est essentielle à la validité, car ceux-ci doivent être reproductibles afin qu'il soit possible de tirer des conclusions quant à la performance des personnes évaluées (Cook & Beckman, 2006). Les *Standards* 2014 mentionnent aussi qu'il existe un lien entre la fidélité des scores à un test et la précision de ces derniers. En effet, une fidélité élevée signifie que les scores sont relativement précis et, donc, que des différences de scores sont plus susceptibles de représenter des différences réelles entre les personnes évaluées. À l'inverse, une fidélité faible signifie que les scores sont plutôt imprécis et que des scores différents risquent de ne pas refléter de différences réelles entre les personnes évaluées.

Plusieurs facteurs peuvent affecter la fidélité. Les *Standards* 2014 mentionnent entre autres la procédure d'évaluation. En effet, si cette procédure change, cela risque d'entraîner des changements dans la fidélité des scores au test. D'autres caractéristiques de l'évaluation peuvent favoriser ou défavoriser la fidélité des scores, comme la clarté des instructions et de l'échelle de notation ou le degré d'accord entre les juges. Le nombre d'items utilisés pour estimer l'habilité ou la performance des personnes évaluées, qu'il s'agisse de questions d'examen, de tâches, de situations cliniques, etc., est également un facteur qui affecte la fidélité. À titre d'exemple, une évaluation de stage reposant sur 20 observations d'un étudiant risque d'être plus fidèle qu'une évaluation reposant sur uniquement deux observations. La fidélité des scores peut aussi différer selon la variabilité de la caractéristique évaluée dans la population étudiée. Plus cette caractéristique varie dans la population, plus la fidélité de l'évaluation risque d'être élevée, puisqu'il est plus facile de différencier entre elles les personnes évaluées. À l'inverse, la fidélité risque de diminuer si la variabilité du construit d'intérêt est faible dans la population étudiée.

Il existe plusieurs méthodes statistiques pour estimer la fidélité des scores. De plus, différentes formes de fidélité peuvent être évaluées, soit la cohérence entre les items du test, la stabilité des scores dans le temps et l'accord entre les juges, aussi appelée fidélité interjuges. Selon le test, il arrive qu'une seule ou que l'ensemble de ces trois formes de fidélité soient évaluées. Quelques-unes des méthodes les plus populaires pour évaluer ces trois formes de fidélité sont présentées dans les paragraphes suivants.

Les méthodes basées sur la cohérence interne ont l'avantage de ne nécessiter qu'une seule administration du test. Elles visent à établir dans quelle mesure les items du test (items, tâches, observations, etc.) produisent des scores reproductibles. Le coefficient alpha de Cronbach (Cronbach, 1951) est le coefficient le plus utilisé dans cette catégorie. Il peut être défini de différentes manières et varie entre 0,00 et 1,00. Son auteur le décrit comme la moyenne de l'ensemble des coefficients de corrélation entre différentes moitiés d'un même test (Cronbach, 1951). Il est également défini comme le pourcentage de variance attribuable aux scores vrais dans la variance totale des scores aux tests (Furr & Bacharach, 2014). Par exemple, un coefficient alpha de Cronbach de 0,70 signifie que 70 % de la variance des scores aux tests est attribuable aux vraies différences entre les personnes et 30 % à de l'erreur aléatoire de mesure. L'équation 1 présente la formule pour calculer ce coefficient. Celui-ci tend à augmenter lorsque le nombre d'observations (p. ex., items) augmente et lorsque la corrélation entre les résultats à ces observations est forte. La littérature donne différents barèmes pour juger du degré de fidélité à partir de la valeur du coefficient alpha de Cronbach. Downing (2004) suggère le barème suivant en contexte d'éducation en pédagogie des sciences de la santé : il faut 0,90 ou plus pour des évaluations à enjeux très élevés (examens de certification), 0,80-0,89 est acceptable pour des évaluations à enjeux modérés (examens de fin d'année), 0,70-0,79 est acceptable pour des évaluations à faibles enjeux (examen dans le cadre d'un cours). Cet auteur ajoute que les évaluations dans le cadre d'un cours ayant un coefficient alpha de Cronbach inférieur à 0,70 peuvent tout de même être utiles si elles sont combinées à d'autres évaluations. De plus, il faut garder en tête que cet indice n'est pas une panacée et qu'il ne permet pas à lui seul de déterminer la qualité des items ou d'une évaluation (Cortina, 1993).

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_t^2}\right)$$

où

n = nombre d'items

σ_i^2 = variance d'un item

σ_t^2 = variance du test

L'estimation de la stabilité temporelle des scores, ou la fidélité test-retest, permet également d'estimer la fidélité des scores à un test. Il s'agit alors de faire la corrélation entre les résultats obtenus par un même groupe de candidats à deux administrations du test. Cette façon de faire se bute toutefois à plusieurs difficultés. D'abord, il est rarement possible d'administrer un même test à deux ou à plusieurs reprises à un groupe de candidats. Ensuite, même si cela était faisable, la connaissance du test par les candidats risquerait d'influencer les résultats à sa deuxième administration. De plus, un changement dans les scores observés à la deuxième administration du test pourrait s'expliquer par un véritable changement dans le score vrai des candidats, surtout si le temps écoulé entre les deux administrations est grand. Il faut donc tenir compte de ces inconvénients dans l'utilisation et dans l'interprétation de la fidélité test-retest.

La fidélité interjuges est une autre façon d'évaluer la fidélité des scores lorsque le test nécessite le recours à plus d'un évaluateur. Il existe trois catégories de méthodes pour l'estimer (Stemler, 2004), soit 1) les méthodes basées sur le consensus entre les évaluateurs, 2) les méthodes basées sur la cohérence entre les évaluateurs et 3) les méthodes basées sur l'application d'un modèle de mesure. Notre description de ces méthodes s'appuie sur celles de Stemler (2004) et de Axelson et Kreiter (2009). Les estimations basées sur le consensus évaluent l'accord exact entre les évaluateurs, observé lorsque ceux-ci ont attribué le même score. Le pourcentage d'accord est une méthode de cette catégorie. Il s'agit de calculer le rapport entre le nombre d'occasions où les deux juges sont en accord et le nombre total d'occasions. L'avantage de cette méthode est qu'elle est simple à calculer et à communiquer. Par contre, elle ne tient pas compte du fait que les deux juges pourraient être d'accord simplement par chance. Afin de tenir compte de cette possibilité, Cohen a développé la statistique kappa (p. ex., Huber et al., 2005). Un kappa de valeur 0 signifie que l'accord entre

les deux juges pourrait être simplement attribuable au hasard. Landis et Koch (1977) suggèrent un barème pour interpréter le degré d'accord selon cette statistique, soit faible (valeurs inférieures à 0,00), léger (entre 0,00 et 0,20), passable (entre 0,21 et 0,40), modéré (entre 0,41 et 60), substantiel (entre 0,61 et 0,80) et presque parfait (entre 0,81 et 1,00). Les méthodes de la seconde catégorie évaluent la cohérence entre les scores des juges. Ces méthodes n'évaluent donc pas si les juges ont un accord exact, mais plutôt s'ils ordonnent les personnes évaluées de la même manière. Les coefficients de corrélation de Pearson, pour les données de niveau de mesure intervalle et distribuées normalement, ou de Spearman, pour des données ordinales ou non normales, peuvent être utilisés pour estimer le degré d'accord entre deux juges. La statistique tau de Kendall (deux évaluateurs) ainsi que le coefficient alpha de Cronbach (plus de deux évaluateurs) peuvent aussi être employés. Il en va de même pour le coefficient de corrélation intraclasse qui, selon la variante choisie, peut servir à évaluer tant le consensus (accord absolu) que la cohérence (accord relatif) (Schubert et al., 1999) entre plusieurs évaluateurs. La troisième catégorie de méthodes renvoie aux analyses statistiques permettant de déterminer l'influence des juges sur la variabilité des scores. Pour ce faire, une analyse de généralisabilité peut être utilisée (Axelson & Kreiter, 2009) ou une analyse à l'aide du modèle à facettes de Rasch (Roberts et al., 2010), aussi appelé modèle multifacettes de Rasch.

La précision des scores

L'estimation de la fidélité des scores permet également de déterminer leur précision. Dans le cadre de la TCT, l'erreur-type de mesure sert d'indicateur de précision. Elle représente la variabilité de l'erreur de mesure et, en ce sens, elle a une valeur de zéro pour le cas hypothétique d'un test dont les résultats seraient parfaitement fidèles (Harvill, 1991). Elle se calcule à partir du coefficient de fidélité et de l'écart-type des scores au test à l'aide de la formule présentée à l'équation 2. Par exemple, dans une situation d'évaluation où le coefficient alpha de Cronbach est de 0,85 et que l'écart-type des scores au test est de 8 points, l'erreur-type de mesure est de $8 \times (1\text{-}0{,}85)^{½} = 3{,}10$. Il est ensuite possible de calculer un intervalle de confiance autour d'un score observé. Dans l'exemple précédent, l'intervalle de confiance à 68 % autour d'un score de 70 serait compris entre $70 \pm 3{,}10$, c'est-à-dire entre 66,90 et 73,10. Pour obtenir un intervalle de confiance à 95 %, il faut alors multiplier cette erreur-type de mesure par 1,96, ce qui donne un total de 6,07. L'intervalle de confiance à 95 % pour le score 70 de

notre exemple est donc compris entre 63,93 et 76,07. Cet intervalle signifie que, chez les candidats ayant obtenu un score de 70, 95 % d'entre eux ont un score vrai se situant entre 63,93 et 76,07.

$$ETM = S_x\sqrt{1 - r_{xx'}}$$

où

ETM = erreur-type de mesure

S_x = écart-type des scores au test

$r_{xx'}$ = coefficient de fidélité

L'analyse de la structure factorielle

L'analyse de la structure factorielle, aussi appelée analyse de la dimensionnalité, cherche à déterminer si le test évalue un seul construit (appelé test unidimensionnel ou unifactoriel) ou plusieurs construits, reliés entre eux ou non (appelé test multidimensionnel ou multifactoriel). La description qui en est faite ici s'appuie sur les ouvrages de Furr et Bacharach (2014) et de Tabachnik et Fidell (2013). Plus spécifiquement, l'analyse de la dimensionnalité d'un test consiste à vérifier à l'aide d'une méthode statistique la présence de sous-groupes d'items plus fortement associés entre eux qu'avec les autres. Ces sous-groupes d'items sont appelés dimensions ou facteurs. Ils peuvent être corrélés entre eux ou non. Voici un exemple pour illustrer ce qu'est la dimensionnalité. Un test pourrait évaluer uniquement la compétence en communication clinicien-patient (test unidimensionnel), alors qu'un autre test évaluant la communication pourrait évaluer la communication clinicien-patient, la communication avec les autres professionnels ainsi que la communication avec les proches des patients (test multidimensionnel). Pour ce test multidimensionnel, les performances des candidats aux différentes dimensions pourraient être corrélées, ouvrant la porte à l'hypothèse d'une compétence générale en communication, ou non, c'est-à-dire qu'il faudrait alors considérer de manière indépendante les performances des candidats à chacune des dimensions.

La dimensionnalité d'un test est importante, car elle a une incidence directe sur le calcul et sur l'interprétation des scores des candidats. Si le test évalue une seule dimension, un score unique reflétant la performance

des candidats par rapport à cette caractéristique est utilisé. Par contre, si le test évalue plusieurs dimensions distinctes, un score est calculé pour chacune de ces dimensions.

Plusieurs méthodes statistiques peuvent servir pour analyser la dimensionnalité d'un test en contexte de TCT. Lorsque l'objectif est d'explorer le nombre de dimensions présentes dans le test, c'est habituellement l'analyse en composantes principales ou l'analyse factorielle exploratoire qui est retenue. S'il existe a priori une ou plusieurs hypothèses claires quant à la structure du test, l'analyse factorielle confirmatoire permet de vérifier dans quelle mesure ces hypothèses sont compatibles avec les données empiriques.

L'analyse d'items

L'analyse d'items est décrite de manière quelque peu différente selon les auteurs. Certains la définissent comme une analyse de la qualité des items qui peut être tant qualitative que quantitative (Anastasi & Urbina, 1997). D'autres la définissent comme l'analyse des caractéristiques statistiques des items (fréquence des options de réponse, moyenne, médiane, écart-type, corrélation item-total, taux de non-réponse, etc.) (Downing, 2009a) ou, de manière plus restrictive, de l'analyse de la difficulté et de la discrimination des items (Haladyna & Rodriguez, 2013). En pratique, l'analyse d'items fait généralement référence à l'analyse des caractéristiques statistiques des items, parfois en se limitant à leur difficulté et à leur discrimination, parfois en s'intéressant à d'autres statistiques. De plus, certaines analyses peuvent être propres à un type d'items en particulier. Notamment, lors de l'examen des questions à choix multiples, une analyse des distracteurs est faite. L'objectif de cette dernière est de s'assurer pour chaque question que tous les choix de réponse sont utilisés, que les distracteurs sont en moyenne choisis par les candidats les moins forts au test et que la bonne réponse est en moyenne choisie par les candidats les plus forts au test.

Dans le contexte de la TCT, l'indice de difficulté, parfois symbolisé par *p* (*p-value*), est la proportion de réussite à un item. Ainsi, pour un item dichotomique (donc pouvant prendre uniquement deux valeurs, 0 = échec et 1 = succès), un indice de difficulté de 0,70 signifie que 70 % des candidats ont réussi l'item. Cet indice peut donc prendre n'importe quelle valeur située entre 0 et 1. Plus l'item est facile, plus son indice de difficulté se rapproche de 1. Pour un item à crédit partiel, pour lequel les candidats peuvent obtenir une partie des points (p. ex., un item sur cinq points où

les candidats peuvent avoir un résultat de 0, 1, 2, 3, 4 ou 5), l'indice de difficulté est sa moyenne. Autrement dit, pour un item sur cinq points, un indice de difficulté de 3,50 signifie que la performance de moyenne à cet item est de 3,50. Il est possible de convertir cette moyenne en proportion (3,50/5 = 0,70) afin d'obtenir un indice de difficulté variant entre 0 et 1 pouvant être comparé à celui d'items dichotomiques.

En ce qui a trait au pouvoir de discrimination d'un item, il décrit la capacité d'un item à mesurer des différences de score total au test entre les candidats évalués (Haladyna & Rodriguez, 2013). Il existe plusieurs indices statistiques pour l'estimer. Celui qui est le plus simple à calculer est l'indice de discrimination *D*, que nous décrivons ci-après en nous inspirant de l'ouvrage de Crocker et Algina (2008). Sa formule est $D = p_e - p_f$, où p_e est la proportion de candidats ayant réussi l'item parmi le groupe des 27 % ayant obtenu les scores les plus élevés au test ; p_f est la proportion de candidats ayant réussi l'item parmi le groupe des 27 % ayant obtenu les scores les plus faibles. Autrement dit, c'est la différence entre l'indice de difficulté estimé pour le groupe supérieur et celui qui est estimé pour le groupe inférieur. Pour qu'un item soit discriminant, il faut que *D* soit positif et, idéalement, supérieur ou égal à 0,20, signifiant que les candidats les plus forts au test sont, en proportion, plus nombreux que les candidats les plus faibles à le réussir. Il faut cependant être prudent et se rappeler que ces barèmes sont présentés à titre de repères et que le choix du seuil minimal dépendra de la nature et du but de l'évaluation. La corrélation item-total corrigée, aussi appelée corrélation item-test, est un autre indice de discrimination. Il est plus complexe à calculer que l'indice *D*, mais les études montrent qu'il est préférable à ce dernier (Crocker & Algina, 2008), sans compter qu'il est désormais implanté dans la plupart des logiciels statistiques. Cet indice est la corrélation entre les résultats à l'item et le résultat à l'ensemble du test, auquel l'item évalué a été retranché. Par exemple, pour un test de dix items, la corrélation item-total corrigée pour l'item 1 est sa corrélation avec le total des neuf autres items du test.

Haladyna et Rodriguez (2013) donnent un barème pour juger la qualité des items à partir des indices de difficulté et de discrimination. Nous résumons ce barème au tableau 1. Ce type de barème aide à éclairer le jugement quant à la qualité des items, mais ne constitue pas une règle absolue. Cependant, il n'est pas rare de voir qu'une valeur de 0,20 soit considérée minimale pour l'indice de discrimination (Everitt & Skrondal, 2010). Par exemple, Loye (2014) présente des balises souvent utilisées pour

l'interprétation de l'indice de discrimination (corrélation item-total corrigée) : moins de 0,10, l'item n'a aucune utilité ; entre 0,10 et 0,19, l'item discrimine trop peu et devrait être révisé ; entre 0,20 et 0,29, l'item discrimine peu ; entre 0,30 et 0,39, l'item discrimine bien ; 0,40 et plus, l'item discrimine très bien. De surcroît, comme l'ajoutent Johnson et Morgan (2016), cet indice aura tendance à être plus faible pour un construit plus large (plusieurs objectifs d'apprentissage) et plus élevé pour un construit plus étroit (un seul ou quelques objectifs d'apprentissage). Il faut également souligner que la difficulté d'un item a une certaine influence sur son pouvoir de discrimination. Un item réussi ou échoué par une très grande proportion de candidats (indice de difficulté de plus de 0,90 ou de moins de 0,10) permet mal de départager les candidats forts des candidats faibles. Dit autrement, si les scores à un item varient peu ou qu'ils ne varient pas, ils ne permettent pas de distinguer les candidats entre eux.

Tableau 1
Barème pour juger la qualité des items

Difficulté	Discrimination	Commentaire
0,60 – 0,90	> 0,15	Item idéal. Difficulté modérée. Item discriminant.
0,60 – 0,90	< 0,15	Item peu ou pas discriminant. À éviter.
> 0,90	Sans importance	Item réussi pas la plupart des candidats. Généralement peu discriminant. À conserver si le contenu couvert par cet item est important.
< 0,60	> 0,15	Item difficile et discriminant.
< 0,60	< 0,15	Item difficile, mais peu ou pas discriminant. Peut être dû à une erreur dans la clé de correction. À éviter ou à réviser.

Note. Adapté de *Developing and validating test item*s par T. M. Haladyna et M. C. Rodriguez, 2013, Routledge.

Avant de terminer cette discussion au sujet de l'analyse des items, il faut aborder l'analyse de la matrice des corrélations inter-items. Cette analyse est parfois escamotée dans les livres sur le sujet, mais elle a son utilité. Elle consiste en l'examen des corrélations entre chacune des paires d'items du test. Par exemple, pour un test de dix items, il faut examiner le sens et la force de la corrélation entre l'item 1 et chacun des autres items, puis, entre l'item 2 et chacun des autres items et ainsi de suite. Les valeurs de ces corrélations nous renseignent sur les relations qu'entretient chacun

des items avec les autres et permettent ensuite de mieux comprendre les indices de discrimination et les résultats de l'analyse de la dimensionnalité. Lorsque les items sont tous censés mesurer le même construit, l'ensemble des corrélations devraient minimalement toutes être positives. Si le test est multidimensionnel, il est possible que certaines paires d'items ne soient pas associées de manière statistiquement significative (corrélations nulles). Toutefois, peu importe la dimensionnalité du test, il ne devrait pas y avoir de paires d'items corrélés négativement, sauf dans la situation exceptionnelle où un test serait composé de facteurs négativement corrélés entre eux. Plus les items d'un test sont corrélés entre eux, plus ils sont cohérents, plus leur discrimination sera bonne (corrélation item-total corrigée) et plus les coefficients de fidélité basés sur la cohérence interne, comme le coefficient alpha de Cronbach, seront élevés. Cependant, les paires d'items pour lesquelles la corrélation est supérieure ou égale à 0,70 méritent d'être examinées pour s'assurer que les items ne sont pas redondants. En effet, une corrélation de cette magnitude signifie que ces items partagent au moins 50 % de variance commune, ce qui laisse penser que leurs contenus sont peut-être trop similaires. Éventuellement, l'un des deux items pourrait être supprimé, en affectant peu l'information recueillie par le test et tout en diminuant sa longueur.

Le fonctionnement différentiel d'item

La vérification de la présence d'items biaisés est une autre preuve de validité liée à la structure interne d'un test. Ce type d'analyse, appelée analyse du fonctionnement différentiel d'items (FDI), est plus fréquent dans le cadre de tests à enjeux élevés. L'objectif de cette analyse est d'évaluer si, à niveau égal d'habileté, des candidats de groupes différents ont des probabilités différentes de réussir l'item (Clauser & Mazor, 1998). Un item est dit non biaisé si, à un niveau d'habileté égal, les candidats de groupes différents ont la même probabilité de réussir l'item. Autrement dit, la probabilité de réussite d'un item ne devrait dépendre que de l'habileté du candidat. Plusieurs variables peuvent servir à regrouper les candidats afin de vérifier la présence de FDI, par exemple le sexe, l'ethnie, la méthode ou l'administration du test (Haladyna & Rodriguez, 2013). Le plus souvent, les analyses de FDI concernent la comparaison entre deux groupes (hommes vs femmes, francophones vs anglophones, etc.), mais certaines méthodes permettent la comparaison entre plusieurs groupes. Il existe différentes méthodes statistiques pour faire une analyse de FDI : la statistique de Mantel-Haenszel, qui est probablement la méthode la plus populaire,

l'analyse de régression logistique, la méthode SIBTEST et les méthodes basées sur la TRI (Haladyna & Rodriguez, 2013). La présentation de ces méthodes dépasse toutefois la portée de cet article.

Les preuves basées sur les relations avec d'autres variables

Il arrive que l'interprétation des résultats à un test implique un lien avec d'autres variables. Par exemple, les tests d'admission aux programmes d'études visent souvent à identifier les candidats les plus aptes à réussir le programme une fois admis. Ainsi, l'interprétation de ces tests a un lien avec la réussite académique future des candidats sélectionnés. Les preuves de validité basées sur les relations avec d'autres variables consistent donc en l'analyse des diverses relations qu'entretient un test avec d'autres variables pertinentes, dans le cas de notre exemple, la réussite académique future.

Les *Standards* 2014 rapportent que les preuves basées sur les relations avec d'autres variables peuvent concerner l'examen : 1) de la convergence et de la divergence avec d'autres tests et 2) de la relation avec différents critères. Évaluer la convergence signifie évaluer l'association entre le test et un autre test mesurant le même construit ou un construit similaire (p. ex., l'association entre les scores à deux tests évaluant le raisonnement clinique). Évaluer la divergence signifie évaluer l'association entre le test et un construit qui devrait y être peu ou pas associé (p. ex., association entre les scores à un test évaluant le raisonnement clinique et les scores à un test évaluant les connaissances sur l'éthique professionnelle). La relation avec un critère concerne l'association entre le test et un critère opérationnellement distinct du test. Dans l'exemple du test d'admission à un programme d'études, nous pourrions vérifier son association avec la moyenne générale au programme, qui servirait de critère. Il est possible d'étudier l'association entre le test et un critère mesuré plus tard dans le temps, c'est-à-dire d'évaluer sa capacité prédictive. Dans cette situation, les données recueillies servent à déterminer si le test permet de prédire le critère. Il est également possible que le critère soit mesuré approximativement en même temps que le test. Dans ce cas, les données recueillies servent à évaluer si le test est associé au critère lorsque les deux sont mesurés environ au même moment. On parle alors de relation concomitante. Une multitude de méthodes peuvent être mobilisées pour documenter les relations entre le test et d'autres variables pertinentes. En effet, il existe une panoplie de méthodes statistiques permettant d'évaluer l'association entre deux ou plusieurs variables (test du khi carré, coefficients de corrélation, régression linéaire ou logistique, systèmes d'équations structurelles, etc.).

Certains facteurs peuvent influencer les résultats obtenus lors de l'analyse des preuves basées sur la relation avec d'autres variables. C'est notamment le cas des degrés de fidélité et de validité avec lesquels les autres variables sont mesurées. L'association entre le test et d'autres variables d'intérêt pourrait être faible parce que les autres variables sont mesurées de manière peu fiable ou peu valide. Dans le cas de l'évaluation du pouvoir prédictif, le temps écoulé entre la passation du test et la mesure du critère à prédire risque d'affecter la force de la relation.

Les preuves basées sur les conséquences du test

L'évaluation des conséquences du test correspond à l'évaluation des effets positifs et négatifs du test. L'idée est que l'utilisation d'un test et l'interprétation de ses résultats a des implications sur plusieurs plans (candidats évalués, apprentissages, programme d'études, enseignants ou superviseurs, patients, société, etc.). Ces conséquences peuvent être attendues ou inattendues, à court ou à long terme, et d'importance variable (Downing & Haladyna, 2009 ; Messick, 1995). Par exemple, ce qui est évalué dans le cadre d'un programme d'études influence nécessairement les efforts d'apprentissage. Dans le cas d'un test à enjeux élevés, comme celui d'un examen de certification, la réussite ou l'échec, ou encore le risque de commettre une erreur en faisant échouer ou réussir un candidat, a des conséquences importantes pour ce dernier, et même, éventuellement, pour la société. Il importe donc d'évaluer si les effets positifs escomptés de l'utilisation du test sont effectivement observés, d'une part, et d'évaluer les effets négatifs et leur ampleur, d'autre part. Ultimement, les avantages de l'utilisation du test doivent l'emporter sur les inconvénients. L'évaluation des conséquences du test comporte donc une part de subjectivité et de jugement de valeur (Downing & Haladyna, 2009 ; Messick, 1989). L'article d'El Hassan (2009) est un exemple d'étude publiée sur l'évaluation de la validité de conséquences d'un test.

Les preuves recueillies concernant les conséquences du test peuvent être variées et dépendent du test, de l'interprétation proposée de ses scores et de la justification de son utilisation (AERA et al., 2014). Pour illustrer ces propos, reprenons l'exemple du test d'admission à un programme d'étude visant à accueillir les candidats les plus aptes à réussir ledit programme. Si les données recueillies appuient cette interprétation des résultats, il sera alors possible d'argumenter que cette conséquence positive du test est démontrée. Les données recueillies pourraient également montrer

que les femmes réussissent mieux que les hommes ce test d'admission. Il faudrait alors démontrer que c'est le cas parce que les femmes évaluées possèdent effectivement davantage la ou les caractéristiques recherchées pour réussir dans ce programme et que le test ne présente pas de biais favorisant les femmes, comme des items pour lesquels la probabilité de réussite, à habileté égale, est supérieure pour les femmes. Dans ce dernier type de situation, il faut démontrer que les différences observées sont dues à de réelles différences sur la caractéristique évaluée et non à un problème de validité du test (AERA et al., 2014 ; Messick, 1995). Lorsqu'un seuil de réussite est utilisé, des données sur la qualité du processus d'établissement du seuil, sur sa crédibilité, sur la précision et sur la fiabilité des scores et du classement des candidats de part et d'autre du seuil pourraient être des éléments permettant d'évaluer les conséquences du test (Downing & Haladyna, 2009).

Enfin, il faut préciser qu'il existe une certaine controverse concernant les preuves de validité basées sur les conséquences de l'évaluation. La plupart des auteurs s'entendent pour affirmer que les conséquences d'une évaluation sont importantes et doivent être prises en considération. Par contre, certains jugent que les conséquences d'une évaluation n'ont rien à voir avec la validité des inférences faites à partir de celle-ci et, conséquemment, que leur étude ne devrait pas faire partie du processus de validation (Linn, 1997 ; Mehrens, 1997 ; Moss, 1998).

Conclusion

La validité est un concept capital en évaluation, puisqu'elle concerne le caractère vraisemblable des interprétations faites à partir des résultats au test. L'évaluation est centrale dans les formations en sciences de la santé. Cependant, les responsables de ces formations, experts dans leur domaine, sont généralement peu familiers avec le concept de validité et avec le processus de validation. L'objectif de cet article était de démystifier ce concept à partir du cadre conceptuel de la validité proposé par les *Standards* 2014 et d'expliquer comment documenter les différents types de preuves de validité dans le contexte de la pédagogie en sciences de la santé.

En pratique, la validation d'une évaluation commence dès sa conception (Mislevy, 2007). Comme le souligne Downing (2006), l'ensemble des étapes de conception d'un outil d'évaluation contribue à justifier l'interprétation faite de ses scores. La conception d'un test nécessite de déterminer

sa finalité (pourquoi), les personnes à évaluer (qui), le contenu à évaluer (quoi), la méthode la plus appropriée (comment) et le moment opportun pour l'évaluation (quand). Comme ces paramètres influencent l'interprétation des scores au test, ils influencent indubitablement le processus de validation, dont le rôle est de vérifier si cette interprétation est justifiée ou si elle ne l'est pas.

Il n'existe pas de règle claire pour déterminer à quel moment mettre fin au processus de validation. Comme le mentionnent les *Standards* 2014, il existe toujours d'autres preuves qui pourraient contribuer à notre compréhension de l'à-propos, de la portée et des limites de l'interprétation des scores à un test. En pratique, ce processus prend généralement fin lorsqu'il y a suffisamment de preuves permettant de soutenir de manière convaincante l'interprétation des scores au test. L'ampleur du processus de validation varie aussi selon l'importance des enjeux liés au test et selon les contraintes d'ordre pratique (ressources humaines, temps et budget disponibles pour cette démarche, etc.). Par ailleurs, les *Standards* 2014 précisent que l'importance des différentes preuves de validité varie selon l'interprétation des scores. Ainsi, dans l'exemple du test d'admission, la démonstration de son pouvoir prédictif, c'est-à-dire de sa capacité à prédire la réussite académique dans le programme d'études, sera incontournable. En revanche, pour une évaluation reposant sur le jugement de plusieurs évaluateurs, les données sur la fiabilité de leur jugement seront essentielles.

Le résultat final du processus de validation est l'argumentaire de validité. Selon les *Standards* 2014, ce dernier « intègre différentes preuves à l'intérieur d'un compte-rendu cohérent concernant le degré auquel les preuves et la théorie appuient l'interprétation suggérée des scores pour une ou plusieurs utilisations spécifiques » (p. 21, traduction libre). Ces preuves peuvent venir d'études nouvelles ou antérieures. De plus, chacune des interprétations et des utilisations des scores au test doit être justifiée par l'argumentaire de validité.

En somme, il n'y a pas de façon unique de procéder à la validation d'un test et ce processus dépend de plusieurs facteurs, tels que l'objectif du test, son importance, l'interprétation souhaitée, le type de test, etc.

Enfin, soulignons que le cadre conceptuel de la validité proposé par les *Standards* depuis 1999 n'est pas sans critiques, ce qui a d'ailleurs mené à la proposition d'autres cadres (Borsboom et al., 2004 ; Kane, 1992, 2006 ; Lissitz & Samuelsen, 2007 ; Newton & Shaw, 2013). Plusieurs auteurs

trouvent que cette conception de la validité est inutilement complexe, difficile à appliquer, insiste trop sur les éléments externes au test, tels que les relations aux autres variables et leurs conséquences, et pas assez sur le contenu et sur le développement du test (Borsboom et al., 2004 ; Lissitz & Samuelsen, 2007). Newton (2016) a un discours similaire à certains égards et propose de parler de micro et de macro-validation pour distinguer les preuves de validité directement liées au test et à son développement (contenu, processus de réponse, qualité des items) de celles associées au test pris globalement et mis en relation avec d'autres variables (pouvoir de prédiction) et suggère que la micro-validation soit considérée essentielle et la macro-validation souhaitable. Finalement, certains vont jusqu'à suggérer l'abandon du concept de validité, étant donné les divergences de points de vue à son sujet dans la littérature, pour celui de qualité (Newton & Shaw, 2013). Le concept de validité risque donc de continuer d'évoluer et de demeurer pour les années à venir un sujet de discussion et de préoccupation tant pour les théoriciens que pour les personnes impliquées dans la validation d'évaluations en sciences de la santé (Kinnear et al., 2024).

Révision linguistique : Marie-Claire Legaré
Mise en page : Emmanuel Gagnon
Résumé en portugais : Eusébio André Machado

Réception : 18 avril 2025
Version finale : 21 novembre 2025
Acceptation : 23 mars 2026

LISTE DES RÉFÉRENCES

Abozaid, H., Park, Y. S., & Tekian, A. (2017). Peer review improves psychometric characteristics of multiple choice questions. *Medical Teacher, 39*(sup1), S50-S54. https://doi.org/10.1080/0142159X.2016.1254743

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7e éd.). Prentice Hall.

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

Axelson, R. D., & Kreiter, C. D. (2009). Rater and occasion impacts on the reliability of pre-admission assessments. *Medical Education, 43*(12), 1198-1202. https://doi.org/10.1111/j.1365-2923.2009.03537.x

Bloom, B. S. (1956). Taxonomy of educational objectives: the classification of educational goals (1e éd.). Longmans, Green.

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

Brennan, R. L. (2006). Perspectives on the evolution and future of educational measurement. Dans R. L. Brennan (dir.), *Educational measurement* (4e éd., p. 1-16). American Council on Education and Praeger.

Clauser, B. E., & Mazor, K. M. (1998). An NCME instructional module on using statistical procedures to identify differentially functioning test items. *ITEMS - Insctructional topics in educational measurement*, 31-44.

Coderre, S., Woloschuk, W., & McLaughlin, K. (2009). Twelve tips for blueprinting. *Medical Teacher, 31*(4), 322-324. https://doi.org/10.1080/01421590802225770

Comité d'agrément des facultés de médecine du Canada. (2023). *Normes d'agrément des programmes d'éducation médicale en vue de l'obtention d'un diplôme en médecine (MD)* https://cacms-cafmc.ca/wp-content/uploads/2023/03/Normes-et-elements-CAFMC-AU-2024-2025.pdf

Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *The American Journal of Medicine, 119*(2), 166.e167-166.e116. https://doi.org/http://dx.doi.org/10.1016/j.amjmed.2005.10.036

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*(1), 98-104.

Crocker, L. M., & Algina, J. (2008). *Introduction to classical and modern test theory*. Cengage Learning.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334. http://dx.doi.org/10.1007/BF02310555

De Champlain, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education, 44*(1), 109-117. https://doi.org/10.1111/j.1365-2923.2009.03425.x

DeVellis, R. F. (2012). Scale development: theory and applications (3e éd.). Sage.

Dore, K. L., Kreuger, S., Ladhani, M., Rolfson, D., Kurtz, D., Kulasegaram, K., Cullimore, A. J., Norman, G. R., Eva, K. W., Bates, S., & Reiter, H. I. (2010). The reliability and acceptability of the Multiple Mini-Interview as a selection instrument for postgraduate admissions. *Academic Medicine, 85*(10), S60-S63 10.1097/ACM.1090b1013e3181ed1442b.

Downing, S. M. (2003). Validity: on the meaningful interpretation of assessment data. *Medical Education, 37*(9), 830-837. https://doi.org/10.1046/j.1365-2923.2003.01594.x

Downing, S. M. (2004). Reliability: on the reproducibility of assessment data. *Medical Education, 38*(9), 1006-1012. https://doi.org/10.1111/j.1365-2929.2004.01932.x

Downing, S. M. (2006). Twelve tips for effective test development. Dans S. M. Downing et T. M. Haladyna (dir.), *Handbook of test development*. L. Erlbaum.

Downing, S. M. (2009a). Statistics of testing. Dans S. M. Downing et R. Yudkowsky (dir.), *Assessment in health professions education* (p. 93-118). Routledge.

Downing, S. M. (2009b). Written assessments: constructed-response and selected-response formats. Dans S. M. Downing et R. Yudkowsky (dir.), *Assessment in health professions education* (p. 149-184). Routledge.

Downing, S. M., & Haladyna, T. M. (2009). Validity and its threats. Dans S. M. Downing et R. Yudkowsky (dir.), *Assessment in health professions education* (p. 21-55). Routledge.

Drennan, J. (2003). Cognitive interviewing: verbal data in the design and pretesting of questionnaires. *Journal of Advanced Nursing, 42*(1), 57-63. https://doi.org/10.1046/j.1365-2648.2003.02579.x

El Hassan, K. (2009, 2009/06/01). Investigating substantive and consequential validity of student ratings of instruction. *Higher Education Research & Development, 28*(3), 319-333. https://doi.org/10.1080/07294360902839917

Everitt, B., & Skrondal, A. (2010). *The Cambridge dictionnary of statistics*. Cambridge University Press.

Fives, H., & DiDonato-Barnes, N. (2013). Classroom test construction: The power of a table of specifications. *Practical Assessment, Research & Evaluation, 18*(3).

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: an introduction* (2e éd.). Sage.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Harvill, L. M. (1991). An NCME instructional module on standard error of measurement. *ITEMS - Instructional topics in educational measurement,* (Summer), 33-41.

Hoban, J. D., Lawson, S. R., Mazmanian, P. E., Best, A. M., & Seibel, H. R. (2005). The Self-Directed Learning Readiness Scale: a factor analysis study. *Medical Education, 39*(4), 370-379. https://doi.org/doi:10.1111/j.1365-2929.2005.02140.x

Huber, P., Baroffio, A., Chamot, E., Herrmann, F., Nendaz, M. R., & Vu, N. V. (2005). Effects of item and rater characteristics on checklist recording: what should we look for? *Medical Education, 39*(8), 852-858. https://doi.org/doi:10.1111/j.1365-2929.2005.02226.x

Jeffrey, D. (2013). *L'éthique dans l'évaluation scolaire*. Presses de l'Université Laval.

Johnson, R. L., & Morgan, G. B. (2016). *Survey scales: a guide to development, analysis, and reporting*. The Guilford Press.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535. https://doi.org/10.1037/0033-2909.112.3.527

Kane, M. T. (2006). Validation. Dans R. L. Brennan (dir.), *Educational measurement* (4e éd., p. 17-64). American Council on Education and Praeger.

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

Kibble, J. D., & Johnson, T. (2011). Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations? *Advances in Physiology Education, 35*(4), 396-401. https://doi.org/10.1152/advan.00062.2011

Kinnear, B., St-Onge, C., Schumacher, D. J., Marceau, M., & Naidu, T. (2024). Validity in the next era of assessment: consequences, social impact, and equity. *Perspectives on Medical Education*. https://doi.org/10.5334/pme.1150

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives: Handbook II: Affective domain*. David McKay Co.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159-174. https://doi.org/10.2307/2529310

Laurier, M., Tousignant, R., & Morissette, D. (2005). *Les principes de la mesure et de l'évaluation des apprentissages* (3e éd.). G. Morin.

Lineberry, M. (2020). Validity and quality. Dans R. Yudkowsky, Y. S. Park & S. M. Downing (dir.), *Assessment in health professions education* (2e éd., p. 17-32). Routledge.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*(2), 14-16. https://doi.org/doi:10.1111/j.1745-3992.1997.tb00587.x

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher, 36*(8), 437-448.

Loye, N. (2014). La conception des examens et des tests : le cas de l'éducation physique. Dans M.-J. Durand & N. Loye (dir.), *L'instrumentation pour l'évaluation : la boîte à outils de l'enseignant évaluateur*. Marcel Didier.

Loye, N. (2018). Et si la validation était plus qu'une suite de procédures techniques? *Mesure et évaluation en éducation, 41*(1), 97-123. https://doi.org/https://doi.org/10.7202/1055898ar

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research November/December, 35*(6), 382-386.

Marceau, M., St-Onge, C., Gallagher, F., & Young, M. (2022). Validity as a social imperative: users' and leaders' perceptions. *Canadian Medical Education Journal, 13*(3), 22-36. https://doi.org/https://doi.org/10.36834/cmej.73518

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practice, 16*(2), 16-18. https://doi.org/doi:10.1111/j.1745-3992.1997.tb00588.x

Messick, S. (1989). Meaning and values in test validation: the science and ethics of assessment. *Educational Researcher, 18*(2), 5-11. https://doi.org/10.2307/1175249

Messick, S. (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Miller, G. E. (1990). The assessment of clinical skills/competence/performance. *Academic Medicine, 65*(9), S63-67. http://journals.lww.com/academicmedicine/Fulltext/1990/09000/The_assessment_of_clinical.45.aspx

Mislevy, R. J. (2007). Validity by design. *Educational Researcher, 36*(8), 463-469. https://doi.org/doi:10.3102/0013189X07311660

Moss, P. A. (1998). The role of consequences in validity theory. *Educational Measurement: Issues and Practice, 17*(2), 6-12. https://doi.org/doi:10.1111/j.1745-3992.1998.tb00826.x

Newton, P. E. (2016). Macro- and micro-validation: Beyond the 'five sources' framework for classifying validation evidence and analysis *Practical Assessment, Research & Evaluation, 21*(12).

Newton, P. E., & Shaw, S. D. (2013). Standards for talking and thinking about validity. *Psychological Methods, 18*(3), 301-319. https://doi.org/10.1037/a0032969

Osterlind, S. J. (2010). Modern measurement: theory, principles, and applications of mental appraisal (2e éd.). Pearson Education/Prentice Hall.

Polit, D. F., & Beck, C. T. (2006). The content validity index: are you sure you know what's being reported? Critique and recommendations. *Research in Nursing and Health, 29*(5), 489-497. https://doi.org/10.1002/nur.20147

Roberts, C., Rothnie, I., Zoanetti, N., & Crossley, J. (2010). Should candidate scores be adjusted for interviewer stringency or leniency in the multiple mini-interview? *Medical Education, 44*(7), 690-698. https://doi.org/10.1111/j.1365-2923.2010.03689.x

Roberts, C., Zoanetti, N., & Rothnie, I. (2009). Validating a multiple mini-interview question bank assessing entry-level reasoning skills in candidates for graduate-entry medicine and dentistry programmes. *Medical Education, 43*(4), 350-359. https://doi.org/10.1111/j.1365-2923.2009.03292.x

Schubert, A., Tetzlaff, J. E., Tan, M., Ryckman, J. V., & Mascha, E. (1999). Consistency, inter-rater reliability, and validity of 441 consecutive mock oral examinations in anesthesiology: Implications for use as a tool for assessment of residents. *Anesthesiology, 91*. https://doi.org/10.1097/00000542-199907000-00037

St-Onge, C., & Young, M. (2015). Evolving conceptualisations of validity: impact on the process and outcome of assessment. *Medical Education, 49*(6), 548-550. https://doi.org/doi:10.1111/medu.12734

St-Onge, C., Young, M., Eva, K. W., & Hodges, B. (2017, 2017/10/01). Validity: one word with a plurality of meanings. *Advances in Health Sciences Education, 22*(4), 853-867. https://doi.org/10.1007/s10459-016-9716-3

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4). https://doi.org/10.7275/96jp-xz07

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6e éd.). Pearson/Allyn & Bacon.

Vice-rectorat aux études et aux affaires étudiantes. (2025). *Règlement des études*. https://www.ulaval.ca/sites/default/files/notre-universite/direction-gouv/Documents_officiels/Reglements/Reglement_des_etudes.pdf

Yang, S.-C., Tsou, M.-Y., Chen, E.-T., Chan, K.-H., & Chang, K.-Y. (2011, 3//). Statistical item analysis of the examination in anesthesiology for medical students using the Rasch model. *Journal of the Chinese Medical Association, 74*(3), 125-129. https://doi.org/10.1016/j.jcma.2011.01.027

Yudkowsky, R., Park, Y. S., & Downing, S. M. (2020). *Assessment in health professions education*. Routledge.