Mesure et évaluation en éducation

Et si la validation était plus qu'une suite de procédures techniques ?

Nathalie Loye

Volume 41, Number 1, 2018

URI: https://id.erudit.org/iderudit/1055898ar
DOI: https://doi.org/10.7202/1055898ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

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

Explore this journal

Cite this article

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/10.7202/1055898ar

Article abstract

Many studies in the field of education focus on the validation of testing instruments, such as questionnaires or assessment instruments. Indeed, such studies are published each year in the journal Mesure et évaluation en éducation. These studies can be classified in two broad categories: on the one hand, those in which psychometric procedures (e.g. factor analysis or Item Response Theory model like the Rasch model) constitute, in and of themselves, the validation process; and on the other hand, those in which relying on an expert panel to develop the instrument is seemingly enough to ensure validity. This article aims to show the necessity of using an argument-based validation process with quantitative and qualitative evidence by proposing two models which can be combined to help formulate arguments supporting the assessment instrument validation. The idea is to address the fact that if we do not understand what we are trying to demonstrate, the arguments are likely to be of little use.

Et si la validation était plus qu'une suite de procédures techniques?

Nathalie Loye
Université de Montréal

Mots clés : validation, conception d'épreuves évaluatives, mesure, arguments de validité

En éducation, de nombreuses études portent sur la validation d'instruments de mesure, tels que des questionnaires ou des instruments d'évaluation. La revue Mesure et évaluation en éducation en publie d'ailleurs chaque année. Ces études peuvent être globalement classées en deux catégories : d'une part, celles dans lesquelles les procédures psychométriques, comme une analyse factorielle ou l'application d'un modèle de la théorie de réponses aux items (p. ex., le modèle de Rasch), constituent en elles-mêmes la démarche de validation ; d'autre part, celles dans lesquelles le recours à un panel d'experts pour concevoir l'instrument semble suffire à garantir sa validité. Cet article tente de montrer qu'il est nécessaire de faire reposer une démarche de validation sur un argumentaire basé sur des preuves de natures quantitative et qualitative, en proposant deux modèles qui, combinés, sont propres à guider la formulation des arguments pour soutenir la validation d'un instrument d'évaluation. L'idée est de contrer le fait que si l'on ne comprend pas ce qu'on essaie de montrer, le risque d'avoir des arguments peu utiles est grand.

Key words: validation, test design, measurement, argument-based validation

Many studies in the field of education focus on the validation of testing instruments, such as questionnaires or assessment instruments. Indeed, such studies are published each year in the journal Mesure et évaluation en éducation. These studies can be classified in two broad categories: on the one hand, those in which psychometric procedures (e.g. factor analysis or Item Response Theory model like the Rasch model) constitute, in and of themselves, the validation process; and on the other hand, those in which relying on an expert panel to develop the instrument is seemingly enough to ensure validity. This article aims to show the necessity of using an argument-based validation process with quantitative and qualitative evidence by proposing two models which can be combined to help formulate argu-

ments supporting the assessment instrument validation. The idea is to address the fact that if we do not understand what we are trying to demonstrate, the arguments are likely to be of little use.

Palavras-chave: validação, conceção de provas de avaliação, medição, argumentos de validade

Em educação, muitos estudos abordam a validação de instrumentos de medição, como questionários ou instrumentos de avaliação, que, de resto, a revista Mesure et évaluation en éducation tem publicado todos os anos. Estes estudos podem ser globalmente classificados em duas categorias: por um lado, aqueles em que os procedimentos psicométricos, como a análise fatorial ou a aplicação de um modelo de teoria da resposta ao item (por exemplo, o modelo de Rasch), constituem em si o processo de validação; por outro lado, aqueles em que o recurso a um painel de especialistas para conceber o instrumento parece ser suficiente para garantir a sua validade. Este artigo tenta mostrar que é necessário basear um processo de validação sobre um argumentário sustentado em evidências de natureza quantitativa e qualitativa, propondo dois modelos que, combinados, são capazes de guiar a formulação de argumentos para apoiar a validação de um instrumento de avaliação. A ideia é contrariar o facto de que, se alguém não entende o que está a tentar mostrar, o risco de ter argumentos pouco úteis é grande.

Note de l'auteure : La correspondance liée à cet article peut être adressée à [nathalie.loye@umontreal.ca].

Introduction

Validité, validation, valider : autant de mots que nous utilisons facilement avec une multitude de sens dans le langage courant. *Valide* peut en effet autant faire référence à la force ou à la bonne santé d'une personne qu'au caractère confirmé d'un événement ou à l'usage approprié d'un objet. Jusque-là, rien à redire : la langue française est ce qu'elle est et chacun la manie à sa guise.

Les choses se compliquent grandement lors de l'utilisation de ces mots dans le langage des sciences humaines et sociales. Que signifie *valider* un questionnaire ? Peut-on dire d'un instrument qu'il est *valide* ? Est-ce que l'épreuve en mathématique que nous avons conçue permet d'obtenir une mesure *valide* des acquis des élèves ? Comment mettre en place une démarche de *validation* ? Autant de questions qui habitent bien des chercheurs et des praticiens en éducation puisque les réponses proposées laissent souvent planer certains doutes – pour ne pas dire des doutes certains – quant à leur réelle valeur.

L'objet de cet article est donc de tenter de mettre un peu d'ordre dans les usages possibles de ces mots lorsqu'ils sont appliqués à un processus de mesure et de réfléchir aux moyens de définir ce qui est valide ou d'établir la façon de valider. Nous n'avons évidemment pas la prétention de régler un problème mille fois discuté et jamais résolu. Nous tenterons plutôt de prendre le temps de s'y arrêter une fois de plus et d'avancer d'un cran dans notre compréhension du sujet.

Dans ce contexte, pas question de parler de validité sans évoquer les débuts de la mesure en psychologie. Gustav Fechner (1801-1887) posait l'hypothèse d'un lien entre la grandeur physique d'un stimulus (p. ex., une lumière plus ou moins brillante, un objet plus ou moins lourd) et la grandeur de la sensation produite sur le système nerveux (Michell, 1999). Même si Fechner n'en parlait pas en termes de validité, au-delà de son besoin d'établir une relation mathématique entre la mesure physique et la mesure mentale, son intérêt était quand même de s'assurer d'avoir un résultat valide pour la dernière en se basant sur la première, en laquelle il avait

confiance puisqu'il avait des instruments (physiques) pour la mesurer. Elizabeth Duffy (1904-1970) s'intéressait aux émotions avec la même préoccupation. Seul le moyen différait puisque la mesure physique proposée était celle de la tension musculaire.

Alfred Binet ne parlait pas non plus de validité (André, Loye et Laurencelle, 2014), mais il considérait que les tests qu'il faisait passer ainsi que les questions et tâches qu'il utilisait permettaient d'établir une mesure de l'intelligence. Ainsi, un pas était franchi de ne pas utiliser une mesure physique intermédiaire comme Fechner et Duffy, mais de concevoir des instruments pour procéder directement à la mesure *de ce qu'il y a dans la tête d'un humain*. Quoi qu'il en soit, la question était quand même, pour Binet, Fechner et Duffy, de présenter des arguments de diverses natures pour soutenir la plausibilité de la mesure obtenue, ce qui est en lien au sens large avec la notion de validité.

C'est justement sur la nature des arguments que se penche cet article, avec l'idée que, si l'on ne comprend pas ce qu'on essaie de montrer, le risque d'avoir des arguments peu utiles ou de mal les utiliser est grand. Après un rapide regard sur l'évolution du concept, nous présenterons et regrouperons deux cadres de référence propres à organiser un argumentaire pour soutenir une démarche de validation d'un instrument d'évaluation. Le modèle de validation de Kane (2006) est d'abord présenté, suivi du modèle *Evidence-Centered Design* de Mislevy et son équipe (Mislevy, Steinberg et Almond, 2003), avec la préoccupation de faire le lien entre les deux. Suit une discussion sur la nature et l'utilité des arguments que la combinaison de ces deux modèles devrait permettre.

Rapide coup d'œil sur l'évolution du concept

André, Loye et Laurencelle (2014) ont proposé un historique de l'évolution du concept de validité, né au début du 20e siècle. Deux constats en émergent. Le premier est qu'il existe diverses formes de validité, toujours évoquées à l'heure actuelle, telles que les validités apparente, concomitante, prédictive ou de contenu, auxquelles s'ajoute la fidélité. Le second est que ces formes de validité ont en général émergé en même temps que les techniques statistiques qui permettent de les étudier. Pensons ici à la corrélation et à la validité en référence à un critère, qu'elle soit concomitante ou prédictive (Guilford, 1946, 1954), à la corrélation et à la validité de contenu (Cureton, 1951) et à l'analyse factorielle et à la validité de

contenu (Cronbach et Meehl, 1955). L'assimilation est telle que, parfois, la démarche de validation est confondue avec les techniques statistiques ou psychométriques, sans autre forme de procès.

À ce titre, Laveault (2012) met en garde contre les mauvais usages de l'alpha de Cronbach (1951) pour étudier la fidélité : le recours à une analyse factorielle est une pratique mille fois présentée dans des études dont l'objet annoncé est la validation d'un instrument (p. ex., Genoud, 2008 ; Harvey, 2012 ; Plante, 2010). Dans d'autres études de validation, c'est toutefois le contenu, sous forme d'un volet qualitatif, qui prend la vedette. La vérification des qualités de la mesure issue de l'instrument n'est alors pas évoquée (p. ex., Demonty, Fagnant et Dupont, 2015 ; Vantourout et Goasdoué, 2014). Dans un cas comme dans l'autre, il est par ailleurs très fréquent qu'aucune définition de la validité ne soit proposée, comme si la démarche choisie allait de soi pour valider.

Vers 1975, la situation est un peu paradoxale. D'une part, un certain consensus existe quant aux caractéristiques nécessaires pour assurer une mesure de qualité en éducation et en psychologie (Newton et Shaw, 2014). En témoigne la version de 1974 des *Standards for educational and psychological tests* (AERA, NCME et APA, 1974), qui généralise une conception de ce qu'est la validité, laquelle est largement adoptée par les praticiens et les chercheurs, tant aux États-Unis que dans de nombreux autres pays. Or, d'autre part, bien qu'étant largement partagée, cette conception de la validité n'est pas pour autant claire et facile à mettre en œuvre, notamment par la juxtaposition de différents types de validité sans vision d'ensemble. Il est déjà question de validité de construit à cette époque, mais son statut est encore flou, même si Loevinger la propose dès 1957 comme ayant le potentiel de réunir tous les types de validité en un tout (Loevinger, 1957).

Il faut attendre la fin des années 1980 pour voir s'imposer la vision moderne de ce qu'est la validité, articulée autour de la notion de validité de construit, mais constituée d'une variété d'ingrédients et d'éléments conceptuels et procéduraux encore disparates. Ce modèle, issu du travail de nombreux chercheurs dont Cronbach, Kane, Linn, Hambleton ou Embretson (voir Newton et Shaw, 2014), se concrétise dans un texte de Messick (1989) qui a fortement marqué la littérature scientifique consacrée à la validité. Messick y discute de manière approfondie de l'insuffisance des diverses formes que sont les validités de contenu et critériée, et prône la nécessité d'une validité unifiée. Celle-ci correspond à la recherche d'une vue

d'ensemble sous la forme d'un réseau nomologique qui met en évidence les relations entre les divers construits en action dans la démarche de mesure avec l'instrument et qui aboutit à l'idée que toute validité est en fait une validité de construit.

Toutefois, il est notoire que ce texte est loin d'être facile à lire et que la validité de construit proposée n'est pas acceptée par tous. La discorde provient en général de la place à accorder à la théorisation du construit à mesurer dans l'établissement de la validité. Par exemple, Lissitz et Samuelson (2007) ont proposé une conceptualisation de la validité qui ne tient pas compte du contexte d'usage du test, ramenant la validité à une propriété du seul test. Dans le même esprit, Borsboom et ses collègues (2009) reprochent à la validité de construit, telle qu'elle est définie dans Messick (1989), le glissement d'une vision ontologique qui suppose, selon eux, de théoriser l'existence même du construit vers une vision épistémologique qui se contente d'une suite de jugements et d'interprétations pour valider l'instrument.

Néanmoins, malgré la complexité et la difficulté à opérationnaliser son modèle de validité unifiée, Messick (1989) le synthétise sous la forme d'une matrice à double entrée. La première concerne l'objectif de la validation, qu'il scinde en séparant l'interprétation de l'usage du test. La seconde concerne la nature des justifications fournies dans le processus de validation. Ces justifications prennent la forme de preuves (*evidence*) liées soit à la pertinence du test en lui-même, soit à des valeurs ou aux conséquences de son utilisation, ce qui introduit les délicates notions d'éthique et d'usage social dans le processus. Ce dernier point est particulièrement objet de controverse, divers auteurs ne considérant pas que ces aspects éthiques et sociaux doivent être inclus dans la validité (voir une discussion à ce propos dans Markus et Borsboom, 2013). De plus, et malheureusement, la ligne de démarcation entre les catégories est loin d'être précise puisque tant les deux types d'objectifs que les deux types de justifications ont largement tendance à se superposer. Le tableau 1 propose toutefois une interprétation de la matrice de Messick, inspirée de McNamara et Roever (2006), qui tente justement de qualifier les différentes cellules de la matrice afin de les différencier.

Tableau 1
Interprétation de la matrice de Messick (1989), inspirée de McNamara et Roever (2006)

	Interprétation : ce que les scores sont supposés signifier	**Usage : ce à quoi les tests servent en réalité**
Preuves liées à la pertinence du test	Quel raisonnement et quels arguments permettent de soutenir les affirmations à propos des performances des candidats ?	Est-ce que ces interprétations ont du sens, sont utiles et justes dans le contexte qui nous intéresse ?
Preuves liées aux valeurs et conséquences	Sur quels présupposés (notamment sociaux et culturels) reposent les construits évalués et les interprétations qu'on fait des scores ?	Qu'est-ce qui arrive (ou pourrait arriver) à notre système éducatif (ou encore plus largement) si l'on utilise ces tests ?

Depuis les années 2000, les écrits sur la validité sont souvent des historiques avec synthèse critique, des débats sur la pertinence de mettre la validité de construit au cœur du processus ainsi que des réflexions tentant de simplifier les manières d'établir la validité. De nouveaux modèles largement inspirés de la validité unifiée ont notamment émergé (p. ex., Kane, 2006 ; Newton et Shaw, 2014). Selon Sireci (2007), même si les méthodes de validation actuelles sont encore bien imparfaites, les démarches basées sur la génération d'arguments obligent à se poser des questions liées à la validité des données produites par les tests et à faire des inférences pour y apporter des réponses, ce qui est déjà très bien.

La section qui suit présente le modèle de Kane, dont l'intention annoncée est de rendre plus simple la démarche lorsque vient le temps de valider. Ce modèle est souvent utilisé pour guider la validation d'épreuves en éducation, probablement parce qu'il met l'accent sur la façon de faire, plutôt que sur la définition de ce qu'est la validité.

Le modèle de Kane

La première idée à la base du modèle de Kane est que la démarche de validation dépend énormément du contexte d'utilisation de l'instrument à valider (Kane, 2009). Lors de la conception d'une épreuve en éducation, par exemple en mathématique, le concepteur ne cherche pas à valider l'épreuve en elle-même. L'enjeu de la démarche de validation est plutôt, selon cette perspective, de savoir si le score attribué représente bien le niveau d'habileté de chaque élève et si la décision prise sur la base de ce score est adaptée. Ce sont donc l'interprétation et l'usage des scores qui ont de l'importance.

Il se peut que l'interprétation soit évidente, par exemple comme c'est le cas en comptabilisant le nombre d'additions réussies parmi une liste par un élève pour établir son score ; il n'y a pas besoin de validation puisque le score correspond simplement à un résumé de la performance de l'élève. Si l'épreuve est conçue dans une visée formative, les enjeux liés aux décisions sont de moindre importance, ce qui minimise en même temps la nécessité d'une validation. Par contre, dès que l'épreuve conçue vise à mesurer un construit plus général – l'habileté en mathématique ou des compétences précises en mathématique –, plusieurs interprétations sont souvent possibles. Par exemple, un score à une épreuve constituée de problèmes en géométrie pourrait être interprété comme un score en géométrie, en résolution de problèmes ou encore en lien avec la vision de l'élève en trois dimensions. Si, en plus, des décisions administratives (admission, certification) ou politiques (réforme du système éducatif) découlent directement des scores obtenus, une démarche de validation est absolument nécessaire.

La différence entre un score correspondant à un nombre d'additions réussies et celui issu d'une épreuve visant à évaluer le niveau d'habileté en mathématique d'un élève est le potentiel à généraliser le résultat. S'il est facile de supposer que l'élève qui a réussi un grand nombre d'additions de la liste sait faire des additions, il est plus que hasardeux de généraliser la note obtenue à une épreuve en mathématique lorsqu'elle combine des exercices et des problèmes de diverses formes et portant sur des contenus variés. C'est donc cette généralisation qui est au cœur de la validation.

Le modèle de validation de Kane est articulé comme une chaîne de quatre inférences, à soutenir par des arguments interprétatifs qui permettent de faire le lien entre :

1) les performances observées (observations) ;
2) la manière d'attribuer un score aux performances (score observé) ;
3) le caractère généralisable du score (score univers) ;
4) la signification du score (score cible) ; et
5) la manière d'utiliser le score pour prendre une décision (Kane, Crooks et Cohen, 1999).

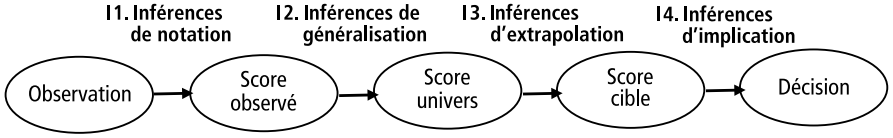

Figure 1 : Les quatre inférences dans le modèle de Kane (adapté de Kane, Crooks et Cohen, 1999)

I1. Les inférences de notation

Elles concernent à la fois les conditions de construction de l'instrument, le design des tâches, la formulation des questions et l'établissement des modalités d'évaluation (incluant la détermination des données qui doivent être collectées dans le processus et la façon dont elles seront jugées). Les inférences de notation sont à considérer au moment de la création de l'instrument et nécessitent de documenter soigneusement la démarche afin de garder traces de chacun des arguments qui permettent de les soutenir.

Cette étape de conception de l'épreuve est complexe, car elle fait intervenir de nombreux éléments, par exemple définir le domaine à évaluer, le type de tâche qui est pertinent, la manière de formuler les items, de les combiner, de les résoudre et de les corriger. Elle fait directement écho à la validité psycho-didactique telle qu'elle est définie par Vantourout et Goasdoué (2014). Pour ces auteurs, la validité repose avant tout sur un état des lieux des connaissances à évaluer ; sur l'étude des variables didactiques qui influencent autant la formulation de la tâche que l'activité de réponse ; ainsi que sur la proximité entre la tâche, la manière dont elle est

présentée à l'élève et les modes d'enseignement habituels auxquels il est soumis. Elle nécessite donc de mettre en place des méthodes de travail rigoureuses et efficaces, et constitue à elle seule un défi de taille.

I2. Les inférences de généralisation

Elles ne prennent place qu'une fois que l'épreuve a fait l'objet d'une étude pilote ou d'un prétest ou qu'elle a été implantée ; c'est-à-dire une fois que des données sur les réponses des candidats sont disponibles. Elles s'intéressent à la stabilité et à la fidélité des scores observés afin qu'ils puissent devenir des scores plus généraux, et non plus liés uniquement à l'épreuve. C'est dans ces inférences que les procédures psychométriques telles que les analyses factorielles, les études de fidélité et de fidélité intercorrecteurs ou intracorrecteurs, les études de généralisabilité, les modèles de la théorie de réponses aux items (TRI) ou encore les études de fonctionnement différentiel des items sont utiles. Dans le cadre de ces inférences, il convient également de définir si des variables peuvent nuire au processus, par exemple si des biais existent dans les tâches ou dans la mise en œuvre, ou encore si les correcteurs utilisent correctement les critères, les grilles ou les barèmes.

I3. Les inférences d'extrapolation

Elles concernent l'étude des sources de variance non souhaitées (p. ex., vérifier si ce n'est pas la compétence en lecture qui affecte le résultat en mathématique). Sur la base des inférences précédentes, elles visent par exemple à s'assurer qu'une tâche permet de discriminer adéquatement en fonction de la formation, qu'elle est réalisable et réaliste, ou qu'elle fonctionne conformément à ce qui avait été prévu au moment de sa conception. À cette étape, l'utilisation d'informations d'autres sources sur les candidats (formation ; dossiers scolaires ; résultats à d'autres épreuves portant sur le même domaine, mais aussi éventuellement issus de mise en œuvre de compétences dans des situations non scolaires) est utile, voire nécessaire pour établir des corrélations ou faire des comparaisons. Ainsi, les inférences d'extrapolation sont à la fois basées sur des preuves de nature analytique en provenance de la définition du domaine à évaluer et de nature empirique.

I4. Les inférences d'implication

Enfin, le rôle des inférences d'implication est, d'une part, de soutenir la crédibilité des interprétations des résultats et des décisions qui en découlent, mais aussi – et surtout – de s'interroger sur les conséquences de ces

décisions. Ces inférences ne sont pas sans rappeler les aspects éthiques et sociaux mis de l'avant dans le modèle de la validité unifiée de Messick, et soulèvent des défis probablement difficiles à relever. Elles donnent lieu notamment à des questions quant au partage des responsabilités entre les concepteurs et les utilisateurs des instruments (Nichols et Williams, 2009) pour les réaliser.

Le modèle de Kane offre l'avantage de présenter une démarche à suivre pour procéder à la validation d'un instrument d'évaluation. Cette démarche est constituée de deux blocs d'inférences distincts. D'abord, les inférences de notation, de généralisation et d'extrapolation forment un premier bloc par leur nature sémantique. Elles visent à définir comment les observations deviennent un score et ce que ce score signifie. Elles reposent sur l'interprétation de descriptions de natures tant qualitative que quantitative, sans qu'un usage spécifique soit pris en considération. Ce bloc concerne la procédure pour concevoir l'instrument d'évaluation et pour s'assurer qu'il permet de se faire une bonne idée de ce que les candidats savent et peuvent faire à partir des traces récoltées. Le second bloc est constitué des inférences d'implication, qui sont de nature politique dans un contexte d'usage précis. Elles sont liées à des interprétations basées sur les décisions qui découlent du processus évaluatif et s'intéressent aux conséquences de ces décisions. Elles seront peu abordées dans ce qui suit.

Dans les deux blocs, force est toutefois de constater que les nombreux écrits de Kane offrent peu de détails sur la mise en œuvre de la production de toutes ces inférences et sur la forme que les arguments peuvent prendre. Pour le bloc 2, l'hypothèse selon laquelle certains contextes et moments sont plus appropriés que d'autres pour réaliser ces inférences semble plausible. Se pose aussi la question de savoir à qui en incombe la responsabilité. En ce qui concerne le bloc 1, Mislevy et ses collègues (Mislevy et Heartel, 2006 ; Mislevy et al., 2003) proposent une chaîne de raisonnement propice à faire le lien avec les inférences du modèle de Kane (McNamara et Roever, 2006). Leur modèle général, intitulé *Evidence-Centered Design* (ECD), offre en effet une structure pour guider et soutenir, à l'aide de preuves (*evidence*), la conception de n'importe quel type de situation d'évaluation des apprentissages. La présentation de ce modèle fait l'objet de la section suivante, dans la perspective de le relier aux trois premières inférences du modèle de Kane : celles du bloc 1.

Le modèle Evidence-Centered Design *(ECD) de Mislevy*

Le modèle ECD est basé sur le principe selon lequel évaluer consiste à fournir des éléments de preuve à partir de ce que les élèves disent, font ou produisent dans quelques circonstances choisies pour inférer ce qu'ils savent ou savent faire de manière générale (Mislevy et al., 2003). Il offre une vision globale de l'évaluation comme résultat d'un argumentaire basé sur des preuves obligatoirement imparfaites, mais qui seront développées le mieux possible. Ce modèle partage donc, avec celui de Kane, l'idée des arguments basés sur des preuves et celle des inférences pour faire le lien entre ce qui est observé ponctuellement et ce qui est ciblé plus largement. Il est donc parfaitement compatible avec les inférences de notation, de généralisation et d'extrapolation. La question est de savoir ce qu'il apporte pour guider l'opérationnalisation de ces inférences et la formulation des arguments qui les soutiennent.

Le modèle ECD est constitué de cinq strates à travers lesquelles il faut circuler, le cas échéant, en effectuant des allers et retours, pour concevoir une épreuve d'évaluation (voir une synthèse à la fin de cette section dans le tableau 2). Les deux premières strates concernent l'analyse et la modélisation du domaine, c'est-à-dire le travail préalable que les experts en mesure, en évaluation et en conception d'items (soit les didacticiens ou les enseignants) doivent obligatoirement réaliser en amont du design d'un instrument d'évaluation, et qui peut éventuellement avoir besoin d'être ajusté, à la lumière du travail effectué dans les strates suivantes. Le fruit de ce travail doit être efficacement présenté – ce qui ajoute une difficulté – de manière à pouvoir y puiser les divers arguments nécessaires à ce qui doit être réalisé dans les trois strates suivantes. Ces dernières concernent la conceptualisation de l'évaluation, son implémentation, puis son implantation.

S1. L'analyse du domaine

Elle consiste à identifier toutes les informations pertinentes et issues de diverses sources à propos du domaine à évaluer. D'abord, il convient de préciser la perspective selon laquelle le domaine est défini. Par exemple, le domaine des mathématiques peut être vu selon une approche behavioriste, cognitiviste ou encore socioculturelle, avec un impact majeur sur le processus de développement (Mislevy et Riconscente, 2005). Doivent être regroupés, triés et évidemment organisés les savoirs, les savoir-faire, les concepts, les théories, la terminologie, éventuellement les différentes

conceptions de ces savoirs, mais également la manière de les représenter, les curricula, les analyses didactiques ou encore les types de situations habituellement utilisées pour mettre en œuvre ces savoirs. Les méthodes pour rassembler ces informations peuvent être variées et inclure par exemple un état des lieux de la recherche scientifique pertinente, l'examen de situations de la vie courante intéressantes à transposer dans des tâches ou encore un sondage auprès de praticiens. Cette analyse devrait aussi permettre une vision du domaine apte à générer des tâches inédites basées sur des arguments solides (Mislevy et Riconscente, 2005).

S2. La modélisation du domaine

Elle consiste à définir les éléments requis pour évaluer dans le domaine qui a été analysé, comment cela devrait être fait et pourquoi, sans toutefois entrer dans les détails opérationnels des tâches et des instruments évaluatifs. Il s'agit ici de baliser l'évaluation afin d'entamer la conception de l'instrument sous ce guidage dans les strates suivantes. Les éléments mis en évidence dans cette strate offrent une base aux inférences de notation du modèle de Kane puisqu'elles concernent le raisonnement sous-jacent à la construction des tâches. Il s'agit en gros de déterminer de manière générale les types d'observations qui permettraient de déduire le niveau d'habileté des candidats et dans quelles sortes de tâches il serait possible de les obtenir. La modélisation du domaine doit rendre explicite la structure de l'argument évaluatif, par exemple sous forme graphique, dans des tableaux ou même dans une application logicielle dédiée (voir plusieurs exemples dans Mislevy et Riconscente, 2005).

S3. La conceptualisation de l'évaluation

Elle constitue évidemment le cœur du processus et a le potentiel de nourrir de manière substantielle les inférences de notation du modèle de Kane. À la suite de la modélisation du domaine, qui explique sous forme narrative de quoi doit être faite l'évaluation, elle permet de mettre en marche la conception des instruments pour évaluer. Mislevy et Riconscente (2005) qualifient cette strate de « machinerie ». Celle-ci est organisée autour de quatre composantes en interaction qui soutiennent l'argument évaluatif : 1) les caractéristiques du candidat, 2) la composante de la tâche, 3) la composante des résultats qui constituent le modèle d'assemblage, et 4) la composante de présentation de la tâche, qui vient le compléter en précisant comment le processus se déroulera concrètement.

- **Les caractéristiques du candidat :** Puisqu'il n'est pas possible de mesurer directement l'habileté des candidats, il convient de proposer des tâches afin que ce qu'ils disent, font ou réalisent pour les résoudre fournisse des informations aptes à caractériser les candidats à l'issue du processus évaluatif. Cette caractérisation correspond à ce qui fournira la mesure. Dans les cas simples, elle prend la forme d'un nombre de bonnes réponses, d'un pourcentage ou d'un score total calculé ou établi de diverses façons possibles, ce qui est globalement en conformité avec la théorie classique des tests. Dans les cas où les habiletés sont plus complexes que des savoirs ou savoir-faire, les modèles psychométriques de la TRI ou ceux à classes latentes sont souvent favorisés. Évidemment, ces choix reposent sur la nature des performances et des tâches et sur des arguments cohérents avec la modélisation du domaine.
- **La composante de la tâche :** Elle correspond aux conditions dans lesquelles les candidats sont placés pour procéder à la mesure, en cohérence avec la modélisation du domaine. Il s'agit de définir tous les éléments qui constituent l'environnement dans lequel le candidat est mis : les types d'énoncés et de questions ainsi que le matériel et les ressources à rendre disponibles. Doivent également être précisés la forme et la nature des productions attendues et des observations ainsi que les critères à prendre en compte pour les décrire.
- **La composante des résultats :** Elle concerne à la fois l'évaluation et la mesure qui en découle. L'évaluation repose sur des critères qui précisent les qualités attendues des productions des candidats pour chaque item. Ces critères peuvent être de nature qualitative, par exemple la pertinence, la justesse ou encore l'efficacité de la stratégie choisie en fonction de ce qui est jugé important au moment de la conception de la tâche. Il s'agit de définir ce qui devra être observé et de quelle manière il faudra le faire. La mesure consiste ensuite à définir comment corriger les réponses des candidats et établir leur score. Cela impose notamment de fournir les clés de réponses des items, de prévoir des grilles d'évaluation ou des rubriques ainsi que de définir de quelle manière la réponse à chaque item contribue au résultat final des candidats, par exemple par le biais de pondérations ou par son lien avec les différents critères.

Le modèle d'assemblage crée une dynamique entre les trois composantes ci-dessus pour définir quelles tâches retenir, leur nombre et leur couverture du domaine, tout en tenant compte de certaines contraintes (p. ex., la disponibilité du matériel informatique, le temps alloué pour l'épreuve). Il n'est pas sans faire penser au tableau de spécification d'un test, même s'il peut prendre des formes variées et inclure plus d'informations. Il permet notamment de s'assurer de l'alignement entre les trois composantes au fil de la construction de l'instrument, plutôt que de proposer une classification des items a posteriori dans un tableau de spécifications (Mislevy, Haertel, Wise Rutstein et Ziker, 2017).

– **La composante de présentation de la tâche :** Elle a pour objectif de préciser comment opérationnaliser le modèle d'assemblage. Y sont traitées les questions relatives au montage de l'épreuve, au mode de présentation des items aux candidats et au matériel à concevoir. Il s'agit de préparer le travail qui sera fait lors de l'implémentation dans la strate suivante en en organisant la logique.

S4. L'implémentation de l'évaluation

Elle fait entrer les concepteurs dans la phase de production du matériel. En conformité avec le travail réalisé dans les strates précédentes, les tâches, les items, les questions, les grilles d'évaluation ainsi que les guides de passation et de correction sont rédigés. Le cas échéant, les algorithmes de correction automatique sont programmés. Même si des ajustements seront obligatoirement nécessaires au moment de l'implantation, le gros du travail de préparation pour la passation est fait dans cette strate. C'est là également que certains items (voire tous) peuvent faire l'objet d'études pilotes ou de prétests afin de voir comment les données alors recueillies s'ajustent aux modèles de mesure choisis précédemment ainsi que de préciser des aspects pratiques (p. ex., la durée de passation). Dès que des données sont recueillies, certaines inférences de généralisation et éventuellement d'extrapolation du modèle de Kane peuvent prendre place ici.

S5. L'implantation de l'évaluation

Elle est la cinquième et dernière strate. Son objectif est de concrètement faire passer l'épreuve et de vérifier comment les candidats interagissent avec les items et les tâches, comment les correcteurs réalisent la correction, comment les résultats sont produits et synthétisés. Cette étape est essentielle pour vérifier si l'épreuve fonctionne en pratique comme attendu

Tableau 2
Parallèle entre le modèle ECD et le modèle de Kane

Modèle ECD			**Modèle de Kane**	
Strates	**Rôle**	**Exemples d'éléments clés**	**Inférences**	**Exemple d'éléments clés**
S1. Analyse du domaine	Regrouper et organiser les informations relatives au domaine à évaluer	– Curriculum, terminologie – Représentations des savoirs – Types de situations – Analyses didactiques – Revue de littérature, enquêtes	Travail préalable aux inférences	
S2. Modélisation du domaine	Préciser les éléments requis pour évaluer le domaine	– Savoirs, savoir-faire, habiletés, processus cognitifs – Types de tâches – Types d'observations à recueillir, liens avec les habiletés	I1. Inférences de **notation**	– Domaine à évaluer – Types de tâches – Types d'observations à recueillir et lien avec les habiletés
S3. Conceptualisation de l'évaluation	Spécifier de quoi sera faite l'évaluation et en organiser la logique	– **Candidat**: scores, mesure – **Tâche**: types d'énoncés et d'items, ressources, nature et qualité des productions – **Résultats**: critères, établissement du score, clés de réponses, pondérations – **Présentation de la tâche**: guide de montage de l'épreuve, liste du matériel à concevoir	I1. Inférences de **notation**	– Formulation des tâches – Productions attendues – Processus de correction – Processus d'établissement du score

S4. Implémentation de l'évaluation	Produire le matériel et faire les ajustements	– Matériel (papier, virtuel, énoncés, outils, grilles, guides de passation et de correction) – Étude pilote/prétest (pour ajuster les formulations, établir et adapter les procédures de passation, ajuster les modèles de mesure)	I1. Inférences de **notation** I2. Inférences de **généralisation** I3. Inférences d'**extrapolation**	– Mise en œuvre pratique de la correction et du processus de notation – Étude pilote/prétest (propriétés métriques des scores, corrélations/ comparaisons sur la base d'autres informations)
S5. Implantation de l'évaluation	Faire passer l'épreuve et étudier les interactions entre les candidats, l'épreuve, les correcteurs et la production des résultats	– Matériel et guide d'accompagnement – Résultats après évaluation – Regroupement des résultats – Présentation des résultats	I2. Inférences de **généralisation** I3. Inférences d'**extrapolation**	– Analyses des données récoltées lors de la passation – Analyses en incluant des données sur les candidats

et pour faire les inévitables ajustements par la suite. La récolte des données permet de réaliser les inférences de généralisation et d'extrapolation du modèle de Kane.

Cette présentation du modèle ECD vise à montrer que ce cadre de référence guide la conception d'instruments d'évaluation de diverses natures avec des visées variées. Tout comme le modèle de Kane, ce modèle s'adapte à une variété de contextes auxquels il s'ajuste en ampleur et complexité. Sa compatibilité avec les inférences de nature descriptive du modèle de Kane est donc, au moins en théorie, facile à mettre en évidence. Le tableau 2 met les deux modèles en perspective. Il servira de base à la réflexion et à la discussion qui suivront sur la nature des arguments tirés du modèle ECD qui pourraient soutenir les inférences de notation, de généralisation et d'extrapolation du modèle de Kane.

Discussion

La nature et la variété des arguments pour soutenir les inférences dans une démarche de validation

Autant Kane que Mislevy et ses collègues réfèrent à la conception de Toulmin (1958) pour définir comment monter un argumentaire. Dans sa forme la plus simple, il est constitué d'une déclaration qui définit ce qui doit être établi, d'hypothèses qui ont le potentiel de la soutenir et de justifications qui explicitent la manière dont elles le font. L'argumentaire peut inclure des hypothèses de différentes natures et des justifications alternatives qui peuvent amener à réfuter la déclaration, lorsque c'est possible, pertinent ou nécessaire. Il faut alors mettre en perspective les diverses justifications pour pouvoir statuer sur la valeur de la déclaration. Réfléchir à la nature et à la variété des arguments pour soutenir les inférences du modèle de Kane nécessite donc de se pencher sur les déclarations, les hypothèses et les justifications qui établissent le lien entre les deux modèles (ceux de Kane et de Mislevy et ses collaborateurs).

L'exercice théorique qui est fait dans cet article nécessitera évidemment plusieurs mises en pratique réfléchies et planifiées, notamment en mathématique, avant d'être en mesure de proposer des exemples concrets et diversifiés d'arguments ainsi formulés en combinant l'usage des deux modèles proposés. En attendant, McNamara et Roever (2006) ont avant nous regroupé les visions de Kane et de Mislevy et ses collègues. Ils consi-

dèrent en effet que ces deux cadres sont complémentaires et permettent de rationaliser les décisions basées sur des tests, tout en étant compatibles avec le domaine des langues tel qu'ils le définissent dans leurs travaux (voir Bachman et Palmer, 1996). C'est d'ailleurs, à notre connaissance, le seul domaine dans lequel un travail de maillage entre ces deux cadres de référence a été réalisé tant de manière théorique (p. ex., Chapelle, 2010, 2011) qu'empirique (p. ex., LaFlair et Staples, 2017).

Pour illustrer comment organiser un argumentaire pour les trois premières inférences du modèle de Kane dans ce contexte d'évaluation en langue, les exemples tirés de Chapelle, Enright et Jamieson (2008/2011) font l'objet du tableau 3. Ils sont le fruit de plus de cinq années de travail de leur équipe pour regrouper patiemment, et après coup, les preuves de validité amassées au fil de l'élaboration du test en anglais langue seconde qu'est le *Test of English as a Foreign Language* (TOEFL) pour les niveaux de compétence de la langue d'enseignement (ETS, 2011).

Ainsi, le tableau 3 présente des exemples concrets de liens qui existent entre les inférences et les strates, et qui produisent les justifications. Par exemple, les inférences de notation (I1) doivent permettre d'affirmer que les tâches du TOEFL aboutissent à des scores qui reflètent bien les savoirs et les habiletés visés. Pour ce faire, il faut évidemment avoir préalablement analysé (S1) et modélisé le domaine (S2), mais aussi avoir fait travailler des experts en conception de critères et de grilles (S3), avoir expérimenté diverses conditions de passation (S4) et avoir analysé les données récoltées pour en vérifier les propriétés métriques (S4 et S5).

De leur côté, les inférences de généralisation (I2) doivent notamment permettre d'assurer la stabilité des scores produits dans les différentes versions du test, ce qui nécessite par exemple d'avoir vérifié cette stabilité de manière empirique dans des études de généralisabilité (S3) ou d'avoir conçu des modèles de tâches en documentant les processus cognitifs que ces tâches requièrent (S2). Elles visent aussi à s'assurer de retrouver dans les données les structures théoriques des construits évalués à l'aide d'analyses factorielles (S3). Les inférences d'extrapolation (I3) doivent apporter des preuves des relations qui existent entre les résultats au TOEFL et d'autres résultats disponibles sur le niveau de compétence linguistique des candidats ou sur leur réussite scolaire (S4).

Tableau 3

Synthèse de l'argumentaire soutenant la validité du TOEFL
(inspiré de Chapelle, Enright et Jamieson, 2008\|2011, traduction libre)

Inférences de Kane	Déclarations	Hypothèses	Justifications	Strates du modèle ECD
	L'observation des performances au TOEFL montre qu'elles sont en lien avec des savoirs et habiletés pertinents dans des situations représentatives du domaine ciblé.	Les types de tâches peuvent être identifiés.	L'analyse du domaine par des chercheurs en linguistique appliquée a permis d'identifier les tâches. Les enseignants et les apprenants ont confirmé l'importance de ces tâches.	S1. Analyse du domaine
		Les savoirs et habiletés peuvent être identifiés.	Les chercheurs en linguistique appliquée ont identifié les savoirs et habiletés nécessaires pour résoudre des tâches.	S2. Modélisation du domaine
		Il est possible d'imaginer des tâches qui requièrent les savoirs et habiletés importants à évaluer.	Des experts ont travaillé de manière systématique pour élaborer des tâches et pour les mettre à l'essai.	
I1. Notation	Les performances dans les tâches du TOEFL sont évaluées afin de fournir des scores qui reflètent bien les savoirs et habiletés ciblés.	Les rubriques pour évaluer les réponses permettent de fournir des informations sur les savoirs et habiletés ciblés.	Les rubriques ont été conçues, testées et révisées par un groupe d'experts qui sont parvenus à un consensus.	
		Les conditions qui encadrent la passation du test permettent de fournir des informations relatives aux savoirs et habiletés ciblés.	Diverses conditions de passation ont fait l'objet d'essais et de révisions, jusqu'à l'obtention d'un consensus par les experts.	
		Les propriétés métriques des items, les mesures produites et le format des tests permettent de prendre des décisions en référence à une norme.	Diverses analyses psychométriques ont été menées au fil des expérimentations, et les ajustements nécessaires ont été faits.	
I2. Généralisation	Les scores observés sont des estimateurs des scores attendus pour les versions parallèles des	Le nombre de tâches dans le test permet d'obtenir des estimations de l'habileté des candidats de manière stable.	Les études de fidélité et de généralisabilité ont permis de déterminer le nombre de tâches requis.	

	tâches et sont stables entre les évaluateurs. Les scores attendus sont en lien avec le construit tel qu'il est défini.	La configuration des tâches permet d'obtenir une mesure interprétable.	Diverses configurations ont été expérimentées pour trouver une configuration adéquate.	S3. Conceptua-lisation de l'évaluation
		Des procédures sont utilisées pour mettre tous les scores sur la même échelle.	Les scores relatifs à l'écoute et à la lecture ont été mis en équivalence (*equating*).	S4. Implémenta-tion de l'évaluation
		Les tâches et leurs spécifications sont bien définies de manière à créer des tâches parallèles et des versions parallèles de test.	Les tâches ont été modelées pour permettre la création de tâches et de tests parallèles.	
		Les savoirs, processus et stratégies nécessaires pour réussir les tâches varient, conformément aux attentes théoriques.	L'analyse des processus mis en œuvre pour réaliser les tâches et leur verbalisation soutiennent le développement des tâches.	S5. Implantation de l'évaluation
		Les performances obtenues dans les nouveaux tests sont reliées à des mesures existantes provenant d'autres tests en langue.	Des corrélations ont été calculées entre les scores du TOEFL et les scores d'autres tests en langue.	
		La structure interne des scores est conforme aux attentes théoriques (dimensions).	Des structures factorielles pertinentes ont été mises en évidence.	
		Les performances observées au test varient en fonction du niveau de compétence linguistique en anglais des candidats.	Des recherches ont été menées et ont mis en évidence de telles relations.	
I3. Extrapolation	La maîtrise de la langue telle qu'elle est évaluée par le TOEFL contribue à assurer la qualité des performances linguistiques dans les milieux de formation.	La performance aux tests est en lien avec d'autres critères relatifs à l'usage de la langue d'enseignement.	Les résultats au TOEFL sont corrélés avec divers autres résultats obtenus par les étudiants dans le cadre de leur formation scolaire.	

Note. Les justifications dans les cases grisées correspondent à des analyses quantitatives; les autres sont qualitatives.

Il ressort de l'examen du tableau 3 plusieurs constats quant à la nature et à la variété des déclarations, des hypothèses et des justifications. D'abord, les déclarations sont relativement générales, mais soulèvent des hypothèses variées et souvent précises qui sont liées tant à la conception (domaine, instruments) qu'aux résultats attendus de diverses modélisations psychométriques des données récoltées. Les justifications s'appuient quant à elles sur les résultats d'analyses très spécifiques, qui peuvent être de nature tantôt qualitative et reposant sur un travail d'experts (p. ex., synthèses, consensus, verbalisations à haute voix), tantôt quantitative pour vérifier si les données obtenues tiennent leur promesse du point de vue de la psychométrie (p. ex., généralisabilité, fidélité, dimensionnalité, TRI) – ou, évidemment, pour guider les ajustements nécessaires afin que ce soit le cas à l'étape suivante.

En outre, le fait que le travail d'organisation de l'argumentaire réalisé par Chapelle et ses collègues (2008/2011) ait été fait a posteriori montre bien le potentiel qu'il y a à combiner ces deux cadres, mais démontre aussi que certaines hypothèses non évoquées ici auraient pu être abordées. Par exemple, aucune information précise n'est fournie quant à la manière d'établir les scores, alors qu'un travail a forcément été réalisé à cet effet.

Nous retenons donc de cet exemple l'importance de procéder à une planification a priori, et évidemment à faire évoluer, regroupant les étapes de conception et d'analyse, en fonction des strates et inférences des deux modèles proposés. Cette planification devrait en outre inclure une procédure de conservation et de classement de toutes les traces nécessaires pour documenter l'argumentaire de validité. Ce seul point pose probablement des défis opérationnels de taille.

Cette planification devrait également tenir compte du contexte du test, au moins pour préciser à quel point la validation est importante et sur quelles inférences il y a plus ou moins lieu de formuler des arguments. En effet, un enseignant qui conçoit une épreuve pour les élèves de sa classe pourrait utiliser l'approche que nous proposons pour la valider, mais sa planification serait beaucoup plus simple que celle d'un organisme externe qui vise à évaluer tous les élèves d'une même année d'études dans un pays, par exemple. Ces deux cas de figure amènent à se poser la question de la place à accorder au processus de mesure. L'enseignant qui conçoit et valide un instrument pour évaluer ses élèves pourrait mettre l'accent sur les inférences de notation et organiser son argumentaire en puisant aux

quatre strates du modèle ECD, sans vraiment se préoccuper de mesure. Au contraire, l'organisme externe qui cherche à mesurer n'aurait pas d'autre choix que de procéder en plus à des inférences de généralisation et d'extrapolation, et même d'implication. Il n'aurait pas d'autre choix que d'inclure des justifications basées sur des analyses psychométriques, en plus probablement d'enrichir l'argumentaire pour soutenir les inférences de notation.

Au début de cet article était abordée l'idée que certains auteurs annoncent un processus de validation basé sur des analyses très spécifiques, parfois qualitatives et parfois quantitatives. Cette manière de faire n'est tout compte fait pas incompatible avec la vision de la validation présentée ici. Elle nécessiterait par contre que ces auteurs circonscrivent ce que leurs analyses permettent d'affirmer quant à la validation de leur instrument. Ce n'est malheureusement pas souvent le cas, le lecteur restant avec l'idée que les résultats ont fait passer l'instrument à l'état de *validé*.

Quoi qu'il en soit, nous considérons à ce stade de notre réflexion que la combinaison du modèle de Kane et du modèle ECD est prometteuse pour guider efficacement la conception et la validation d'épreuves en mathématique ou dans toute autre discipline.

Conclusion

Cet article est le fruit d'une réflexion sur la démarche de validation qui puise à la fois dans la dure confrontation à la réalité de la conception d'une épreuve en mathématique (voir Loye et Lambert-Chan, 2016) et dans une littérature abondante et rarement facile à lire sur la validité. Le modeste exercice proposé ici vise à montrer que la validation d'une épreuve se réalise et se réfléchit en même temps que sa conception. Il vise également à attirer l'attention sur le fait que la validation est bien éloignée de procédures techniques à appliquer, mais repose au contraire sur une bonne compréhension de ce qui est à vérifier afin d'apporter des preuves utiles pour construire une démarche de validation et prendre conscience de ce qu'il est possible d'affirmer. Le double cadre de référence proposé ici nous semble à cet effet plein de promesses et nous prévoyons l'expérimenter rapidement. Cet article nous a permis d'avancer d'un cran dans notre compréhension du processus de validation, et nous espérons qu'il en sera de même pour le lecteur.

Réception : 5 septembre 2017

Version finale : 30 janvier 2018

Acceptation : 22 février 2018

RÉFÉRENCES

AERA, NCME, & APA (1974). *Standards for educational and psychological tests*. Washington, DC: AER.

André, N., Loye, N. et Laurencelle, L. (2014). La validité psychométrique : un regard global sur le concept centenaire, sa genèse et ses avatars. *Mesure et évaluation en éducation*, *37*(3), 125-148. doi: 10.7202/1036330ar

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford, UK: Oxford University Press.

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Scholten, A. Z., & Franic, S. (2009). The end of construct validity. In R. W. Lissitz (Ed.), *The concept of validity: Revision, new directions and applications* (pp. 135-170). Charlotte, NC: Information Age Publishing.

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008/2011). *Building a validity argument for the Test of English as a Foreign Language*. New York, NY: Routledge.

Chapelle, C. A. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, *29*(1), 3-13. doi: 10.1111/j.1745-3992.2009.00165.x

Chapelle, C. A. (2011). Validity argument for language assessment: The framework is simple... *Language Testing*, *29*(1), 19-27. doi: 10.1177/0265532211417211

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *12*, 1-16. doi: 10.1007/BF02310555

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

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational Measurement* (1st ed., pp. 621-694). Washinton, DC: American Council on Education.

Demonty, I., Fagnant, A. et Dupont, V. (2015). Analyse d'un outil d'évaluation en mathématiques : entre une logique de compétences et une logique de contenu. *Mesure et évaluation en éducation*, *38*(2), 1-29. doi: 10.7202/1036761ar

Educational Testing Service [ETS] (2011). Validity evidence supported the interpretation and use of TOEFLiBT Scores, *TOEFLiBT Research Insight*, series I, vol. 4. Princetown, NJ: ETS. Retrieved from www.ets.org/s/toefl/pdf/toefl_ibt_insight_s1v4.pdf

Genoud, P. (2008). Validation d'un instrument mesurant le climat d'études perçu par les étudiants universitaires. *Mesure et évaluation en éducation*, *31*(1), 31-49. doi : 10.7202/1025012ar

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

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York, NY: McGraw-Hill.

Harvey, L. (2012). Évaluation des compétences dans un programme de formation en enseignement : validité de construit curriculaire. *Mesure et évaluation en éducation*, *35*(2), 69-95. doi: 10.7202/1024721ar

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed., chap. 2). Westport, CT: American Council on Education and Praeger.

Kane, M. (2009). Validating the interpretations and uses of test scores. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, new directions and applications* (pp. 39-64). Charlotte, NC: IAP.

Kane, M., Crooks, T., & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5-17. doi: 10.1111/j.1745-3992.1999.tb00010.x

LaFlair, G. T., & Staples, S. (2017). Using corpus linguistics to examine the extrapolation inference in the validity argument for a high-stakes speaking assessment. *Language Testing*, *34*(4), 451-475. doi: 10.1177/0265532217713951

Laveault, D. (2012). Soixante ans de bons et mauvais usages du alpha de Cronbach. *Mesure et évaluation en éducation*, *35*(2). doi: 10.7202/1024716ar

Lissitz, R. W., & Samuelson, K. (2007). A suggested change in terminology and emphasis regarding validity in education. *Educational Researcher*, *36*(8), 437-448. doi: 10.3102/0013189X07311286

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports*, Monograph supplement 9, *3*, 635-694. doi: 10.2466/pr0.1957.3.3.635

Loye, N. et Lambert-Chan, J. (2016). Au cœur du développement d'une épreuve en mathématique dotée d'un potentiel diagnostique. *Mesure et évaluation en éducation*, *39*(3). doi : 10.7202/1040136ar

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

McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Malden, MA: Blackwell Publishing.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-104). Washington, DC: American Council on Education/Macmillan.

Michell, J. (1999). Measurement in psychology: Critical history of a methodological concept. Cambridge, UK: Cambridge University Press.

Mislevy, R. J., & Heartel, G. D. (2006). Implications of Evidence-Centered Design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 8-20. doi: 10.1111/j.1745-3992.2006.00075.x

Mislevy, R. J., Haertel, G., Wise Rutstein, D., & Ziker, C. (2017). *Assessing model-based reasoning using Evidence-Centered Design: A suite of research-based design patterns*. Cham, Switzerland: Springer.

Mislevy, R. J., & Riconscente, M. M. (2005). *Evidence-centered assessment design: Layers, structures, and terminology*. Menlo Park, CA: SRI International.

Mislevy, R. J., Steinberg, L., & Almond, R. G. (2003). On the structure of educational assessments. *Interdisciplinary Research and Perspectives*, *1*(1), 3-62. doi: 10.1207/S15366359MEA0101_02

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational and psychological assessment*. Thousand Oaks, CA: SAGE Publications.

Nichols, P. D., & Williams, N. (2009). Consequences of test score use as validity evidence: Roles and responsibilities. *Educational Measurement: Issues and Practice*, *28*(1), 3-9. doi: 10.1111/j.1745-3992.2009.01132.x

Plante, I. (2010). Adaptation et validation d'instruments de mesure des stéréotypes de genre en mathématiques et en français. *Mesure et évaluation en éducation*, *33*(2), 1-34. doi: 10.7202/1024894ar

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher*, *36*(8), 477-481. doi: 10.3102/0013189X07311609

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, UK: Cambridge University Press.

Vantourout, M. et Goasdoué, R. (2014). Approches et validité psycho-didactiques des évaluations. *Éducation et formation*, e-302. Repéré à https://halshs.archives-ouvertes.fr/halshs-01239551/document