Mesure et évaluation en éducation

Les enjeux en évaluation des compétences langagières

Michel Laurier

Volume 44, Number 3, 2021

L'évaluation des compétences langagières : enjeux et perspectives

URI: https://id.erudit.org/iderudit/1093064ar
DOI: https://doi.org/10.7202/1093064ar

See table of contents

Publisher(s)

ADMEE-Canada

ISSN

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

Explore this journal

Cite this article

Laurier, M. (2021). Les enjeux en évaluation des compétences langagières. *Mesure et évaluation en éducation*, *44*(3), 5–28. https://doi.org/10.7202/1093064ar

Article abstract

While emerging, the field of language assessment focused on adults learning English. We now observe an enlarged vision that blurs the distinction between first and second languages and considers learners diversity. Tracking this evolution, six major issues can be identified, each one represented by an abundant literature: 1) the nature of language competence and its components, 2) the challenges of the need for authenticity, 3) the complexity of the validation process, 4) the ethical considerations to guide the designers and users, 5) social expectations linked to language evaluation and, 6) the promises of technology applications for language assessment. As a result of this analysis, it appears that the field can be examined from several perspectives – psychometric, linguistic, pedagogical, and social.

Les enjeux en évaluation des compétences langagières

Michel Laurier
Faculté d'éducation, Université d'Ottawa

Mots-clés : évaluation de la langue, compétence langagière, authenticité, validation, éthique, aspects sociaux, technologie

À son émergence, le domaine de l'évaluation des langues focalise sur l'apprentissage de l'anglais chez les adultes. Aujourd'hui, une vision élargie s'impose où la frontière entre langue maternelle et langue seconde finit par s'estomper et où la diversité des apprenants est prise en considération. En suivant cette évolution, six grands enjeux se dégagent, autour desquels existe une abondante littérature : 1) la nature de la compétence langagière et ses composantes, 2) les défis de la recherche de l'authenticité, 3) la complexité du processus de validation, 4) les considérations éthiques qui devraient guider les concepteurs et les utilisateurs, 5) les attentes sociales liées à l'évaluation des langues et, 6) les voies qu'ouvre l'utilisation des technologies pour évaluer les compétences langagières. Cette analyse montre que le domaine peut être examiné d'une perspective psychométrique, linguistique, pédagogique ou sociale.

Key words: language assessment, language competence. authenticity, validation, ethics, social aspects, technology

While emerging, the field of language assessment focused on adults learning English. We now observe an enlarged vision that blurs the distinction between first and second languages and considers learners diversity. Tracking this evolution, six major issues can be identified, each one represented by an abundant literature: 1) the nature of language competence and its components, 2) the challenges of the need for authenticity, 3) the complexity of the validation process, 4) the ethical considerations to guide the designers and users, 5) social expectations linked to language evaluation and, 6) the promises of technology applications for language assessment. As a result of this analysis, it appears that the field can be examined from several perspectives – psychometric, linguistic, pedagogical, and social.

Palavras-chave: Avaliação linguística, competência linguística, autenticidade, validação, ética, aspetos sociais, tecnologia

À medida que foi emergindo, o campo da avaliação linguística concentrou-se na aprendizagem de inglês em adultos. Hoje, vemos impor-se uma visão mais ampla onde a fronteira entre língua materna e segunda língua acaba por se esvair e onde a diversidade dos aprendentes é levada em consideração. Ao acompanhar esta evolução, podemos identificar seis grandes questões em torno das quais encontramos uma literatura abundante: a natureza da competência linguística e seus componentes, os desafios da busca pela autenticidade, a complexidade do processo de validação, as considerações éticas que devem orientar os concetores e os utilizadores, as expectativas sociais relacionadas à avaliação linguística e, em última análise, os caminhos abertos pelo uso da tecnologia para avaliar as competências linguísticas. A partir desta análise, percebe-se que o domínio pode ser examinado sob uma perspetiva psicométrica, linguística, pedagógica ou social.

Introduction

Au début des années 1970, à la faveur d'un engouement pour l'apprentissage de l'anglais comme langue véhiculaire, naît aux États-Unis un nouveau champ d'études, le *Language Testing*. L'émergence de ce champ d'études s'explique par les besoins de concevoir des tests de langue qui permettent une mesure présentant diverses preuves de validité, fidèle et pratique de la maîtrise de l'anglais comme langue seconde pour, par exemple, mener des activités commerciales, participer à des réseaux de recherche, poursuivre des études ou voyager à travers le monde. Compte tenu des besoins, le champ s'est d'abord construit autour de l'évaluation de l'anglais chez les adultes. Les premiers travaux poursuivent alors la voie tracée par Lado (1961) et s'inscrivent dans le courant des approches structurales américaines par la suite nommées les « méthodes audio-orales » (Germain, 1993). Ce champ bénéficie au départ de la popularité aux États-Unis du *Test of English as a Foreign Language* (TOEFL), dont la première administration date de 1964 alors que la *Modern Language Association* en est responsable (Spolsky, 1995). L'année suivante, le TOEFL est pris en charge par le *College Board* avant de passer sous l'égide des *Educational Testing Services* (ETS). Fidèles aux principes de la linguistique contrastive, les concepteurs privilégient alors l'évaluation d'éléments discrets de la langue. L'approche connaît d'autant plus de succès que les approches psychométriques alors en vogue favorisent la mesure d'éléments décontextualisés. Le TOEFL permet l'essor d'autres tests de langue standardisés à grande échelle comme le *International English Language Test* (IELTS) dans le monde britannique. Tous ces instruments ont évolué selon les courants qui ont traversé le domaine de la didactique des langues et plus spécifiquement le champ alors nouveau du *Language Testing*. Du côté francophone, le développement d'instruments d'évaluation à grande échelle s'est fait un peu en marge de ce champ. Pensons notamment au *Diplôme d'études en langue française* (DELF) auquel est associé le *Diplôme approfondi de langue française* (DALF).

L'évaluation en langue a connu les remises en question qui ont affecté le domaine plus large de la mesure et de l'évaluation des apprentissages. Il faut d'ailleurs noter que la revue scientifique la plus en vue dans le domaine de l'évaluation des langues, *Language Testing*, dont le premier numéro remonte à 1984, dispute le statut de revue phare depuis la naissance en 2004 de la revue *Language Assessment Quaterly* dont le titre reflète un changement majeur de paradigme. Ainsi, la tradition du testing, qui renvoie à des évaluations externes souvent à grande échelle et ayant fait l'objet d'une standardisation, peut prêter moins d'attention aux pratiques essentielles d'évaluation qui ont lieu dans le cadre de la classe (*classroom assessment*) dans une perspective de régulation des processus d'apprentissage et d'ajustement des stratégies d'enseignement. Dans le monde francophone, la littérature abondante (Bonniol & Vial, 1997 ; Scallon, 2000) autour de l'évaluation formative témoigne de ce changement qui a touché également la didactique des langues. On reconnaît maintenant l'importance de dispositifs que mettent en place les enseignants et qui s'intègrent dans la réalité de la classe afin de favoriser les apprentissages. Toutefois, les épreuves à grande échelle n'ont pas disparu pour autant.

L'évolution du domaine de la didactique des langues a mené à la transformation des approches en évaluation de la langue. Ainsi, la naissance des approches communicatives a été déterminante. Sous l'influence de la sociolinguistique qui met l'accent sur la réalisation d'actes langagiers (Hymes, 1974 ; Widdowson, 1978), la tendance à se limiter à l'évaluation d'éléments isolés de la langue a été remise en question. De plus, l'importance de fournir aux candidats des situations authentiques, c'est-à-dire des situations comparables à celles auxquelles ils sont susceptibles d'être confrontés dans l'usage de la langue cible, a été mise de l'avant. Par ailleurs, l'accent est dorénavant mis sur les processus cognitifs sollicités dans l'utilisation d'une langue en contexte. Il faut également souligner que l'analyse des contextes d'utilisation effective de la langue a fait ressortir la fragilité de la distinction entre la langue première et les langues additionnelles (seconde, tierce ou étrangère), même si l'on ne parle plus de langue maternelle. En effet, les situations de langues en contact qui se multiplient avec les échanges commerciaux, les communications à l'échelle mondiale ou l'immigration engendrent une complexité où la distinction apparaît fragile (Schissel et al., 2019 ; Thomas & Osment, 2020). Il devient ainsi difficile de concevoir une didactique des langues vivantes qui se distingue de la didactique de la langue première.

En suivant l'évolution du champ que constitue l'évaluation des langues, certains enjeux majeurs apparaissent de façon récurrente (Aryadoust et al., 2021) et six d'entre eux retiennent notre attention dans la suite. Ce sont : 1) la nature de la compétence langagière et ses composantes, 2) les défis de la recherche de l'authenticité, 3) la complexité du processus de validation, 4) les considérations éthiques qui devraient guider les concepteurs et les utilisateurs, 5) les attentes sociales liées à l'évaluation des langues et 6) les voies qu'ouvre l'utilisation des technologies pour évaluer les compétences langagières.

La nature de la compétence

Selon Tardif (2006), une compétence se définit comme « un savoir-agir complexe prenant appui sur la mobilisation et la combinaison efficaces d'une variété de ressources internes et externes à l'intérieur d'une famille de situations » (p. 22). Ces ressources peuvent être des connaissances, des habiletés ou des attitudes de sorte qu'elles ne devraient pas se limiter à des éléments d'ordre cognitif, mais pourraient inclure des éléments d'ordre affectif. Il reste que ce sont plus les composantes cognitives que les composantes affectives qui ont retenu l'attention. Cette définition largement partagée est l'aboutissement d'une réflexion sur le concept de compétence où l'apport de la linguistique a été déterminant. Chomsky (1965) définit la compétence comme un ensemble de règles internalisées et l'oppose à la performance, qui est la manifestation physique et contextualisée de la mise en œuvre de la compétence. Hymes (1974) élargit le concept en y greffant les aspects sociolinguistiques de façon à ne pas limiter la compétence langagière à sa composante cognitive et linguistique. La distinction entre la compétence et la performance est fondamentale dans le domaine de l'évaluation des langues parce qu'elle pose clairement la compétence comme un construit inobservable dont il faut induire la mise en place par une analyse de la performance. Il importe alors de faire produire une performance qui permette effectivement d'établir le lien entre ce qui est observé et la compétence à évaluer. Ce lien logique n'est pas toujours facile à établir, particulièrement au moment de l'évaluation des habiletés réceptives (écoute et lecture) où la performance observable est relativement limitée ; dans ces situations, il faut parfois inférer la compétence à partir de mesures indirectes.

L'expression « compétence linguistique » est un raccourci commode qui ne rend pas compte des mécanismes qui sous-tendent une performance. D'une part, on voit que les aspects linguistiques, c'est-à-dire la morphosyntaxe, la graphie/phonologie et le vocabulaire, ne permettent pas de l'expliquer entièrement et qu'il faut aussi considérer les aspects discursifs, sociolinguistiques et pragmatiques. D'autre part, la question se pose à savoir s'il faut parler de compétence en lecture et de compétence à l'écrit tout en parlant de compétence en interaction orale, compte tenu du fait que l'expression et la compréhension sont rarement dissociées dans les situations de communication orale les plus courantes. Il est probablement plus juste de parler de compétence langagière en reconnaissant le fait que cette compétence intègre des ressources linguistiques, sociolinguistiques et pragmatiques, lesquelles se mobilisent de façon différente selon que la communication est orale ou écrite et qu'elle engage la personne comme émettrice ou comme réceptrice. Il faut ensuite se demander si cette compétence langagière peut être divisée en sous-compétences.

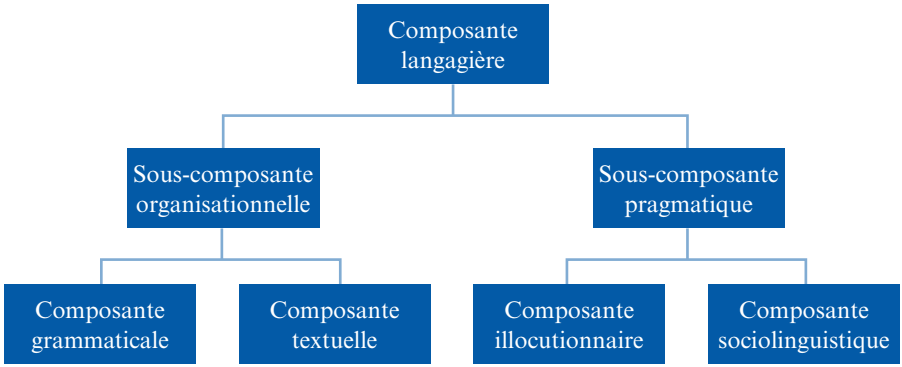
Dans son ouvrage de 1980, Carroll rapporte qu'un examen des corrélations entre différents tests de performance l'a amené à conclure qu'un facteur général peut rendre compte de 58 % de la variance des scores. Ce constat s'inscrit dans un débat ravivé par les travaux d'Oller (1979) qui associe ce facteur général à une « grammaire de l'expectative » et défend l'hypothèse du trait unitaire. L'application de cette grammaire de l'expectative permettrait d'anticiper les éléments du discours et de surmonter les obstacles qui peuvent entraver la communication. S'appuyant sur cette hypothèse, Oller soutient qu'un moyen de mesurer la grammaire de l'expectative est de recourir au test de closure. Par la suite, certaines recherches ont effectivement montré des corrélations notables entre des épreuves de grammaire ou de vocabulaire (Hanania & Shikhani, 1986) et des tests de closure, laissant ainsi entendre que ce type de test pouvait témoigner de la maîtrise générale de la langue malgré les réserves déjà exprimées quant à leur valeur intégrative et à la possibilité de capter des processus de haut niveau (Alderson, 1980). Utilisant une procédure d'analyse factorielle semblable à celle d'Oller pour comparer 22 mesures différentes, Scholz et al. (1980) ont appuyé l'hypothèse du trait unitaire. Cette hypothèse a cependant vite été contestée, notamment par Carroll (1983) qui a repris les données pour démontrer que le facteur général est souvent un artefact de la méthode d'analyse. Hulstijn (1985) a ensuite confirmé l'insuffisance

de l'analyse factorielle et prôné l'utilisation d'un modèle qui intègre différentes facettes (grammaire, vocabulaire, aspects culturels, par exemple) en interaction les unes avec les autres.

La proposition de Hulstijn (1985) rappelle la distinction entre l'approche psychométrique et l'approche édumétrique (Carver, 1974 ; Phakiti & Isaacs, 2021). Sous l'angle psychométrique, la compétence langagière apparaît comme une compétence complexe dont les composantes sont interreliées au point où les analyses de dimensionnalité font souvent émerger un facteur dominant (Blais & Laurier, 1995 ; Fouly et al., 1990). Sous l'angle édumétrique, il apparaît souvent plus utile de distinguer les différentes composantes de la compétence langagière. À cet égard, le modèle proposé par Bachman (1990) au début des années 1970 réconcilie les deux approches et permet de rendre compte de la complexité de la compétence langagière. Il faut d'abord noter que Bachman place la compétence stratégique au cœur de la communication humaine. Cette compétence stratégique, qui est intégrée dans la compétence communicative dans le modèle de Canale et Swain (1980) dont s'inspire Bachman, est maintenant au cœur de la communication humaine, car elle regroupe l'ensemble des ressources servant à gérer les interactions. Dans cette perspective, la compétence langagière est tributaire de la compétence stratégique. Celle-ci fait agir les qualités personnelles qui contribuent à une communication efficace. Elle permet aussi la coordination des mécanismes psychophysiologiques associés au langage, l'intégration des connaissances qu'a acquises le locuteur sur l'objet de la communication (et, plus largement, sur le monde) et, évidemment, le déploiement de la compétence langagière. Une évaluation centrée sur la compétence langagière n'a donc pas à inclure tous les éléments qui relèvent de la compétence stratégique.

Ainsi que l'illustre la Figure 1, Bachman distingue dans la compétence langagière, une sous-compétence organisationnelle et une sous-compétence pragmatique. La première inclut une composante grammaticale au sens large (c'est-à-dire le vocabulaire, la morphosyntaxe et la graphie/phonologie) ainsi qu'une composante textuelle qui permet l'organisation des unités plus larges du discours afin d'en assurer la cohérence. La seconde inclut une composante illocutionnaire qui permet la réalisation effective et efficace des actes langagiers et une composante sociolinguistique qui se rapporte aux règles sociales d'usage et aux références culturelles. Les

Figure 1
La compétence langagière selon Bachman (1990)



Source : Bachman, 1990

sous-composantes ne se développent pas toutes au même rythme. Par exemple, en début d'apprentissage, l'attention se concentrera sur les éléments de base de la composante grammaticale.

L'authenticité

Bachman (1990) formule également quelques remarques à propos de l'authenticité en signalant que cette dernière est liée tant aux situations qui servent à contextualiser une tâche évaluative qu'aux éléments de compétence qui sont effectivement mobilisés dans sa réalisation. La notion d'authenticité est fondamentale dans le domaine de l'évaluation des langues. Elle s'impose, d'une part, avec la préoccupation née de l'approche communicative en didactique des langues en ce qui a trait à l'utilisation pédagogique de documents réels plutôt que fabriqués (Gilmore, 2007) et à la mise en place de situations d'apprentissage apparentées aux situations d'utilisation effective de la langue (Viswanathan et al., 2018). Le défi de l'enseignement est donc d'outiller l'apprenant pour qu'il affronte la complexité d'une situation de communication authentique avec des moyens qui, en début d'apprentissage, sont limités. D'autre part, l'authenticité s'inscrit dans le développement du courant de l'évaluation authentique (*Authentic Assessment*), issue des travaux de Wiggins (1989), qui propose de soumettre à l'apprenant des tâches évaluatives complexes où sont reproduites des situations qu'il peut rencontrer en dehors du cadre scolaire (Koh, 2014).

Comme l'accent est mis sur l'authenticité, celle-ci devient un élément à considérer dans la validité d'une tâche évaluative. Une telle tâche doit s'intégrer dans une situation qui s'apparente à une situation de communication réelle au sein de laquelle les différentes composantes de la compétence interagissent. Cependant, cette exigence d'authenticité est-elle un gage de validité ? En premier lieu, il faut mentionner que la performance d'un locuteur dans une situation authentique est conditionnée par un grand nombre d'éléments qui ne sont pas associés à la compétence et qui peuvent se confondre avec elle de telle manière que le résultat devient difficilement interprétable. De plus, même lorsque ces éléments qui ne font pas partie du construit jouent un rôle assez limité, le caractère intégratif d'une tâche authentique fait en sorte que le résultat reste l'expression de l'interaction de plusieurs éléments de la compétence qui ne sont pas faciles à isoler. Si cela contribue à la validité dans une situation d'évaluation sommative, cela augmente le risque de fournir peu de pistes d'intervention dans le cadre d'une évaluation formative. Il est en effet difficile de diagnostiquer les difficultés d'un élève à partir d'un tel résultat, même en recourant à une grille d'évaluation qui identifie les éléments à observer.

Il faut aussi signaler qu'une situation de communication authentique comporte souvent des éléments qui la rendent plus ou moins complexe selon les caractéristiques des personnes qui y participent. Pensons d'abord à la familiarité avec un contenu particulier qui peut contribuer à faciliter un échange. Pensons également à des biais qui rendent une tâche plus difficile pour certains groupes d'individus en raison de contenus réservés à d'autres groupes, de références culturelles qui ne sont pas partagées ou de la présence d'éléments qui déclenchent une réaction affective pouvant perturber un candidat ou un élève. Historiquement, la détection des biais s'est concentrée autour de tâches susceptibles de produire des résultats différents selon l'origine ethnique ou le genre. Toutefois, dans plusieurs tests de langue destinés aux adultes, le biais associé au fait que la proximité de la langue maternelle avec la langue cible et le niveau de scolarisation favorisent la réussite à une tâche est connu. Certes, des questions se posent sur les effets d'une éradication des biais sur le construit d'un test de langue, surtout s'il prend en compte les éléments de la sous-compétence pragmatique. En revanche, la détection de biais ne doit pas être découragée, car elle permet d'identifier des contenus et des tâches qui favorisent indûment un groupe. Le défi est donc de détecter les éléments qui ne sont pas pertinents et qui engendrent ainsi des problèmes d'équité. C'est dans cet esprit

que Zumbo (2007) propose, dans le but d'éviter une erreur systématique engendrée par une tâche qui favoriserait ou défavoriserait un groupe, une combinaison de méthodes d'analyse du comportement différencié des items (DIF) qu'il faut compléter par une analyse de contenu.

Il est clair que la notion d'authenticité a facilité le développement d'épreuves de langue pour des domaines d'activité spécifiques en offrant la possibilité d'appuyer leur validité. Par exemple, s'il s'agit du processus de vérification de la validité d'une épreuve conçue pour évaluer la capacité de professionnels de la santé à exercer leurs fonctions dans un langue seconde, la vraisemblance des tâches en regard des situations de communication dans le cadre de la prestation de soins de santé permet de s'assurer que l'épreuve induit la performance attendue (Laurier et al., 2021).

Le processus de validation

Le processus de validation est présenté comme une démarche visant à montrer qu'un test possède le degré de validité attendu (Anastasi, 1986). Ce processus soulève toutefois des questions dans son opérationnalisation. D'abord, il est clair qu'il faut non seulement démontrer la validité du test mais aussi sa fidélité, si tant est que ces deux attributs peuvent clairement se distinguer. De fait, depuis les travaux de Messick (1989), la validité de construit est au centre du concept de validité en même temps que ce concept est élargi de manière à couvrir la fidélité et même, comme nous le verrons plus loin, les conséquences de l'utilisation d'un test (Shepard, 1993). La recherche de la fidélité vise à minimiser la variance des scores qui dépend de l'erreur de mesure et est, de ce fait, une condition de la validité. En revanche, cette dernière suppose que la variance dépend des caractéristiques du construit. Il faut cependant souligner que les définitions mêmes du construit peuvent différer. Par exemple, Simon (2011) compare les définitions de la lecture de trois épreuves standardisées utilisées en Ontario pour montrer que ce construit n'est pas nécessairement univoque.

L'une des visions les plus fécondes et les plus originales des dernières années consiste à voir la validation comme un processus argumentatif (Loye, 2018). La validité n'est jamais définitive et se démontre par l'accumulation des preuves qui tendent à montrer qu'un instrument évalue effectivement ce qu'il doit évaluer dans les conditions où il devrait être utilisé. En d'autres termes, la validation doit établir que les diverses inférences que fera l'utilisateur d'un test seront justifiées. C'est l'approche préconisée par Kane (2006, 2012) qui, par rapport à l'approche de Messick (1989), se

veut plus pragmatique dans la mesure où il ne s'agit pas nécessairement de démontrer la présence sous-jacente d'un construit, mais plutôt de défendre la chaine des inférences sur laquelle s'appuie l'interprétation. Kane (2012) propose un modèle où il distingue cinq niveaux d'inférence. Le passage d'un niveau à l'autre implique un type d'inférence particulier.

- La notation – Les observations devraient produire des scores qui reflètent les éléments de la compétence dans le domaine d'utilisation ciblé.
- La généralisation – Les scores observés devraient correspondre à ce que l'on attend dans d'autres situations semblables.
- L'extrapolation – Les scores doivent rendre compte des éléments de la compétence (le construit) qui sont mis en œuvre pour réaliser la tâche.
- L'implication – Les éléments de la compétence (le construit) déterminent la qualité de la performance effective dans le domaine.
- L'utilisation – Les scores conduisent à des décisions qui sont conformes à la fonction prévue de l'évaluation et qui engendrent des effets positifs.

Des méthodologies appliquant les principes de la validation de Kane (2012) ont été développées pour la validation d'épreuves standardisées à grande échelle portant sur la compréhension orale (Aryadoust, 2013) ou mettant l'accent sur les aspects pragmatiques (Youn, 2015) ou encore, utilisant des grilles d'appréciation (Knoch & Chapelle, 2018). La variété des types de preuves qui doivent être invoquées pour démontrer la justesse des inférences qui s'opèrent aux différents niveaux du modèle de Kane (2012) impliquent l'utilisation d'approches tant quantitatives que qualitatives. On privilégie donc une approche mixte pour établir la chaine d'inférences. La robustesse et le nombre des arguments dépendent des enjeux auxquels est liée la décision qui doit éventuellement être prise (Cook et al., 2015). La valeur de certaines inférences n'est cependant pas toujours facile à établir, particulièrement lorsqu'il y a des problèmes d'observabilité comme c'est le cas pour l'évaluation des habiletés réceptives. De plus, pour les niveaux supérieurs d'inférence, le processus d'inférence est plus difficile et il faut davantage tenir compte du contexte d'utilisation de la langue (Bachman, 2005). Par exemple, s'il est relativement facile de démontrer la correspondance des scores d'une épreuve de vocabulaire par une analyse du contenu de l'épreuve, il est beaucoup plus difficile de démontrer jusqu'à quel point

il est possible de généraliser le résultat. Dans cette perspective, il faut voir la validation comme un processus continu qui exploite des données recueillies de diverses sources et qui doit être revu à mesure qu'évoluent les contenus, la nature des tâches ou l'utilisation des résultats.

Les considérations éthiques

Comme nous l'avons remarqué, le débat autour du processus de validation est lié à l'évolution du concept de validité. Depuis Messick (1980, 1989), les conséquences sociales, intentionnelles ou non, sont des éléments à prendre en considération dans l'utilisation d'un test pour des fins déterminées. Si la validité d'un test dépend dans un premier temps de ce que le test est censé mesurer, cette validité peut être remise en question lorsque le résultat de la mesure n'est pas utilisé pour les fins pour lesquelles le test a été conçu. L'idée de Messick d'examiner la correspondance entre les inférences réalisées et l'usage d'un instrument de mesure s'est vite répandue (Moss, 1992). Elle a amené des chercheurs et des concepteurs de tests à parler de « validité des conséquences » même si plusieurs se montrent hésitants à intégrer cette préoccupation dans le concept même de validité (Cizek, 2012 ; Mehrens, 1997). Que les conséquences soient associées à la validité ou non, le débat autour de la question a fait ressortir l'importance, tant pour les concepteurs que pour les utilisateurs, de s'interroger sur les valeurs qui sous-tendent un test, de même que sur son rôle dans le façonnement des valeurs sociales dominantes. Ce questionnement s'inscrit dans le développement d'une éthique autour des effets de l'évaluation.

Dans cette perspective, il convient d'abord de prêter attention aux stratégies que les répondants mettent en œuvre pour réussir un test de langue, particulièrement lorsque celui-ci est une épreuve à enjeux critiques, c'est-à-dire une épreuve dont les résultats risquent d'avoir des conséquences significatives sur l'avenir des répondants. Hamp-Lyons (1997) décrit comme un effet de reflux (*washback*) le phénomène par lequel l'utilisation d'un test détermine ce qui est important et provoque ainsi des modifications dans la finalité des stratégies d'apprentissage et dans la nature des interventions pédagogiques. Du côté des stratégies d'apprentissage, l'effet se manifeste surtout par diverses formes de bachotage qui amènent des élèves à user de stratégies pour réussir le test plutôt que pour véritablement apprendre la langue. Faisant suite aux observations de Nevo (1989), selon lequel les stratégies que mettent en œuvre des candidats pour réussir un test de langue ne reflètent pas toujours les processus mentaux qui sont déployés

dans une utilisation normale de la langue, Wall et Alderson (1993) ont montré les effets pervers de l'effet de *washback*. Du côté des interventions pédagogiques, l'effet se manifeste par une tendance à enseigner en fonction du test et, ainsi, à aligner le contenu des programmes sur le contenu des épreuves, ce qui finit par constituer une menace pour la validité. Cheng et Curtis (2004) rappellent que, bien que ce soit surtout les aspects négatifs de l'effet de *washback* qui sont retenus, celui-ci peut comporter des aspects positifs. Le défi sur le plan éthique serait donc de minimiser les aspects négatifs de l'effet de *washback* et de tirer profit de ses aspects positifs.

Les aspects négatifs peuvent inclure des comportements associés à la tricherie et qui entrent en conflit avec les principes moraux partagés au sein d'une société. Ainsi, devant l'ampleur du phénomène de substitution de personnes au moment de la passation de plusieurs tests de langue à enjeux critiques, Fulcher (2011) va jusqu'à dire qu'il faut éviter cette forme d'évaluation pour appuyer certaines décisions, en donnant comme exemple la sélection d'immigrants.

Il est étonnant de constater que l'effet de *washback* peut présenter une certaine forme de récursivité de sorte que les pratiques reconnues ou tout au moins courantes en évaluation finissent par s'ériger en modèle et perpétuent des pratiques qui soulèvent des questions sur le plan éthique. Dans cette perspective, il est important que les concepteurs et les utilisateurs des instruments qui servent à l'évaluation se dotent de principes qui peuvent les guider.

Le *Joint Committee on Standards for Educational Evaluation* (JCSEE) a été formé en 1975 afin de diffuser des normes de pratique en ce qui a trait à l'évaluation aux États-Unis et au Canada. Cet organisme fait la promotion de normes d'éthique pour l'évaluation des apprentissages en classe (Klinger et al., 2015). Pour ce qui est plus spécifiquement de l'évaluation des compétences langagières, *l'International Association of Language Testing Association* (ILTA) a d'abord publié un code d'éthique (ILTA, 2000) qui réunit une série de neuf principes que les professionnels du domaine devraient suivre :

- Respecter les personnes évaluées ;
- Utiliser de l'information obtenue avec discernement ;
- Adhérer aux règles éthiques des milieux ;
- Utiliser à bon escient sa compétence professionnelle à évaluer ;

– Mettre à jour et partager sa compétence professionnelle ;
– Respecter la profession ;
– Faire la promotion d'une évaluation responsable ;
– Assumer ses obligations sociales ;
– Refuser d'intervenir si les risques d'effets négatifs sont grands.

Ces principes ont ensuite été articulés dans des lignes directrices que l'organisme rend disponibles afin de baliser la pratique de l'évaluation des langues sur le plan déontologique (ILTA, 2007).

Les attentes sociales liées à l'évaluation

Comme on peut le voir dans les principes que propose l'ILTA, beaucoup d'enjeux éthiques sont liés à la dimension sociale de l'évaluation. Pour McNamara et Roever (2006), un bon test sur le plan psychométrique n'est pas nécessairement un bon test sur le plan social. Ces auteurs décrivent le rôle social des tests de langue, notamment comme instruments pour faciliter la reddition de compte dans les systèmes éducatifs, pour permettre le contrôle des flux migratoires et pour définir des groupes sociaux. De fait, en s'éloignant de la fonction formative de l'évaluation, laquelle est principalement motivée par l'objectif d'aider les élèves à mieux apprendre, il est possible d'observer le fait que les instruments qui servent à l'évaluation répondent souvent à des besoins sociaux. Pensons aux tests de langue utilisés dans le cadre de l'admission des étudiants dans les établissements d'enseignement supérieur ou à la sélection des personnes qui désirent immigrer dans un pays ; ces instruments contribuent à une forme d'exclusion sociale. On peut aussi penser aux évaluations de type sommatif comme celles qui servent à attester l'atteinte des objectifs d'un programme scolaire ou celles qui servent à certifier qu'un candidat est apte à exercer certaines tâches professionnelles dans une autre langue ; dans ces cas, à la dimension pédagogique, se greffe une demande sociale plus ou moins explicite.

McNamara (2006) considère que les modèles de validation issus des travaux de Messick (1989) offrent des réponses inadéquates aux questions que soulève l'évaluation en lien avec les valeurs et les contextes sociaux. Le rôle social des tests de langue confère aux concepteurs et aux utilisateurs de tests un pouvoir dont il est facile d'abuser. Shohamy (2001) examine différents tests de langue utilisés pour montrer comment ils s'imbriquent dans le tissu social et peuvent devenir des outils servant diverses visées politiques. Ce pouvoir des tests de langue est d'autant plus préoccupant

que leurs concepteurs et leurs utilisateurs subissent des pressions dont la source remonte à des représentations naïves de la nature et du fonctionnement d'une épreuve de langue. Pour s'en convaincre, il suffit de penser aux débats récurrents sur les bulletins scolaires où la note apparaît comme une finalité incontournable entourée d'une aura de scientificité masquant un processus d'évaluation plutôt opaque.

La mise en place d'une épreuve de langue paraît souvent comme une solution miraculeuse à un problème social ; cette solution émane généralement d'un rapport de force inégalitaire. Le développement d'une épreuve, quel que soit son usage, semble être un processus simple permettant d'offrir des réponses incontestables qui ne sont pas sujettes à interprétation. Les épreuves de langue, quant à elles, se heurtent, d'une part, à des représentations liées à la mesure et, d'autre part, des représentations liées à la langue, ce qui nécessite une double entreprise de déconstruction. Tant du côté de la mesure que du côté de la langue, il est intéressant de noter la persistance d'approches qualifiées, à partir de points de vue très différents, de « normatives ».

L'approche normative de la mesure s'oppose à l'approche critériée. Cette distinction remonte à Glaser (1963). L'approche normative consiste à interpréter les résultats qu'obtiennent les candidats à une épreuve à partir des résultats calculés pour le groupe de référence. Cette approche est pourtant inadéquate quand il faut répondre à des attentes sociales qui s'expriment régulièrement comme l'attestation de la compétence langagière, le rehaussement des exigences dans la maîtrise de la langue ou encore le soutien aux élèves en difficulté. Dans ces cas, c'est plutôt une approche critériée qu'il faudrait privilégier puisque le résultat prend du sens lors de l'analyse de l'écart de la performance observée par rapport au niveau de performance souhaité, plutôt qu'en analysant l'écart avec la performance moyenne d'un groupe de référence. L'approche normative conduit souvent à attribuer un rang aux élèves. C'est souvent celle qui prévaut dans l'imaginaire collectif et, conséquemment, celle que certains voudraient voir dominer dans les tests de langue.

Selon Legendre (2000), une norme linguistique se définit comme un « ensemble de recommandations déterminées par une partie de la société et précisant ce qui doit être reconnu parmi les usages d'une langue afin d'obtenir un certain idéal esthétique ou socioculturel » (p. 903). Une approche normative de la langue est considérée comme un système fermé

qui s'impose au locuteur lorsque celui-ci doit utiliser un registre de langue soutenu. Poussée à l'extrême, cette approche, qui se veut prescriptive, peut réduire la compétence langagière à la capacité à s'exprimer « sans fautes ». L'importance accordée au respect des règles et des conventions dictées par la norme peut d'ailleurs varier, mais les concepteurs de tests de langue doivent en tenir compte pour assurer l'acceptabilité sociale des instruments qu'ils proposent. Par exemple, Laurier et Baker (2015) ont montré que, pour vérifier la maîtrise de la langue par les enseignants et les enseignantes du Québec, quand la maîtrise de l'anglais comme langue d'enseignement est évaluée, les attentes quant au respect de la norme sont moins contraignantes que celles du français

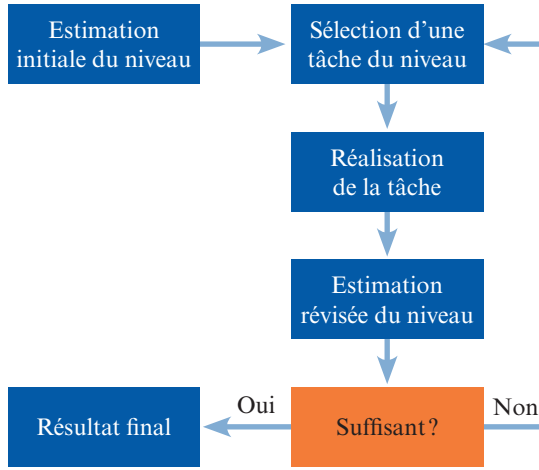
L'utilisation des technologies

L'une des attentes sociales à l'égard de l'évaluation des langues est sans doute l'utilisation accrue des technologies. Les avantages habituellement reconnus à l'ordinateur et aux technologies qui en dérivent justifient cette attente : traitements numériques complexes, branchements en cours d'exécution, intégration de divers types d'information (images, textes, son…) et réduction des contraintes de distance.

Avant les années 2000, l'utilisation des tests adaptatifs suscitait beaucoup d'espoir (Brown, 1997). L'élaboration d'un test adaptatif suppose d'abord la mise en place d'une ou de plusieurs banques d'items qui mesurent un attribut commun. Chaque item est calibré afin de lui associer un certain nombre de paramètres dont le plus important est son indice de difficulté. Cet indice sert à retrouver dans la banque l'item qui est le plus approprié, compte tenu du niveau du candidat. Ce niveau est ajusté après chaque nouvelle réponse. Il en résulte une épreuve qui cible davantage le niveau de compétence, ce qui, par rapport à un test traditionnel, rend la passation plus courte pour un niveau de fidélité égal et plus agréable pour le candidat (Laurier 1992, 1999). Le diagramme de la Figure 2 illustre le déroulement d'un test adaptatif simple.

Une décennie plus tard, la possibilité d'utiliser des techniques de traitement des langues naturelles et de reconnaissance de la parole s'ajoute à l'adaptabilité (Chapelle & Chung, 2010). Cependant, plusieurs problèmes ne sont toujours pas résolus aujourd'hui. Suvorov et Hegelheimer (2013) rappellent que la plupart de tests adaptatifs sont élaborés en appliquant des modèles psychométriques issus de la théorie de réponses aux items. Ces modèles postulent que les tâches respectent le principe

Figure 2
Déroulement d'un test adaptatif



d'unidimensionnalité, c'est-à-dire qu'elles peuvent s'aligner sur l'axe de développement d'un attribut unique. Or, comme nous l'avons vu, la compétence langagière est plutôt multidimensionnelle. De plus, la recherche d'authenticité conduit à des tâches intégratives qui ne satisfont pas toujours l'exigence d'unidimensionnalité.

Par ailleurs, il aurait été souhaitable que les performances des élèves dans les sections qui font appel à des habiletés productives (expression orale et écriture) puissent être corrigées de façon automatisée. Malgré des progrès notables en ce sens, la correction automatisée de productions écrites et orales en anglais dans le but d'inférer les éléments de la compétence langagière ne s'avère pas à la hauteur, notamment parce qu'elle peut être déjouée par certaines stratégies de passation et qu'elle se concentre sur les phénomènes de surface (par exemple, l'orthographe). Les problèmes d'unidimensionnalité et les lacunes de la correction automatisée expliquent pourquoi plusieurs firmes de tests à grande échelle qui avaient développé des épreuves de type adaptatif ou tenté d'automatiser la correction sont revenues vers des méthodes plus traditionnelles.

Isbell et Kremmel (2020) examinent comment sept grands fournisseurs de tests de langue ont raffiné leur procédure de passation à domicile dans le contexte de la pandémie de COVID 19 qui complique les passations sur place. La plupart de ces tests évaluent la maîtrise de l'anglais, mais certains sont offerts dans d'autres langues. Cependant, la passation à distance de tests à enjeux critiques pose des défis importants, particulièrement en ce qui a trait à la vérification de l'identité des candidats et à la divulgation des contenus de l'épreuve. Plusieurs dispositifs combinent la surveillance humaine à distance avec la détection et l'analyse de comportements suspects à l'aide de techniques d'intelligence artificielle. Ces techniques permettent d'entrevoir des moyens d'accroître l'accessibilité et la flexibilité des passations. Papageorgiou et Manna (2021) affirment qu'au-delà des considérations liées à l'utilisation d'outils ou au recours à d'autres personnes, il faut s'interroger sur la comparabilité de ce mode de passation par rapport à d'autres modes qui devraient être équivalents, sur l'équité, sur la protection de la vie privée et sur l'importance de minimiser l'exposition du contenu des épreuves.

Chapelle et Douglas (2006) soulignent que l'innovation est une façon de penser l'évaluation qui émerge une fois dépassée la simple recherche d'efficacité. Il y a une décennie, Chalhoub-Deville (2010) affirmait que l'avenir de l'utilisation des technologies pour l'évaluation de la compétence langagière allait dépendre de la capacité des concepteurs à innover plutôt qu'à reproduire les pratiques antérieures. La pandémie de COVID-19 semble avoir beaucoup fait avancer le domaine à cet égard.

Contrairement à ce que nous pourrions croire, le principal avantage des technologies pourrait ne pas résider dans les applications liées aux tests à grande échelle conçus pour des fins sommatives. Ainsi, les élèves apprécient particulièrement la flexibilité et la variété que permet un environnement virtuel dans le cadre d'activités d'évaluation formative (Milliner & Barr, 2020). Chapelle et Voss (2019) font d'ailleurs remarquer qu'un espace d'innovation, en ce qui a trait à l'exploitation des technologies dans le domaine de l'évaluation des langues, est la possibilité d'offrir aux élèves plus d'occasions d'apprendre à travers les activités évaluatives, car la technologie permet d'individualiser ces activités, de les placer dans un cadre plus stimulant que celui de la salle de classe et d'élargir la gamme des processus d'apprentissage. Par exemple, le logiciel *Abracadabra,* qui permet de diagnostiquer des difficultés en lecture dans un environnement

ludique, illustre de quelle manière un dispositif d'évaluation peut servir l'apprentissage (Abrami et al., 2015). Dans cette situation, la cible d'évaluation coïncide avec la cible d'apprentissage dans l'activité pédagogique.

Conclusion

Cette analyse des principaux enjeux en évaluation des langues confirme le caractère multidisciplinaire du domaine. L'analyse fait ressortir les différentes perspectives à partir desquelles il faut examiner le développement et l'utilisation des tests de langue.

- La perspective psychométrique – Cette perspective se concentre sur les aspects liés à la mesure. Les questions qui se posent sont les suivantes : Est-ce que le construit est unidimensionnel ? Comment le jugement peut-il être modélisé ? Quelles méthodes faut-il employer pour la validation ?
- La perspective linguistique – L'accent est mis sur la nature de la compétence langagière et sur la congruence des tâches avec l'objet à évaluer. Les questions portent donc sur les composantes du construit et sur la manière de construire des tâches qui soient en lien avec ce construit.
- La perspective pédagogique – De ce point de vue, l'intérêt porte davantage sur le rôle de l'évaluation dans l'enseignement et dans l'apprentissage. L'évaluation peut prendre diverses formes, mais la visée formative devient prioritaire. L'une des préoccupations importantes est de savoir de quelle manière améliorer la rétroaction afin de favoriser l'apprentissage.
- La perspective sociale – L'accent est mis ici sur la dimension sociale de l'évaluation de sorte que les pratiques évaluatives s'insèrent dans un contexte social. Il faut alors pouvoir répondre à une demande sociale en maximisant les effets positifs et en minimisant les effets négatifs.

Ces perspectives se complètent mais peuvent aussi s'opposer. Ainsi, d'une perspective psychométrique, les compétences langagières peuvent être décrites comme des construits relativement unidimensionnels alors que la perspective linguistique distingue plusieurs composantes. De même, l'exigence d'authenticité s'impose, pour des considérations différentes selon que la perspective est linguistique ou pédagogique, mais elle soulève

certains problèmes d'une perspective psychométrique. Nous pouvons aussi constater que la perspective sociale fait émerger des défis qui pourraient être ignorés si nous nous en tenions seulement aux autres perspectives. Ajoutons à ces exemples, le fait que l'utilisation des technologies ne présente pas le même intérêt selon une perspective psychométrique ou pédagogique ou même sociale. C'est pour ces raisons que cette analyse des enjeux actuels de l'évaluation des compétences langagières se voulait une invitation à jeter un regard pluriel sur les théories et sur les pratiques qui traversent le domaine.

Réception : 3 décembre 2021
Version finale : 30 mai 2022
Acceptation : 31 mai 2022

LISTE DES RÉFÉRENCES

Abrami, P., Borohkovski, E., & Lysenko, L. (2015). The effects of ABRACADABRA on reading outcomes : A meta-analysis of applied field research. *Journal of Interactive Learning Research, 26*(4), 337-367. Association for the Advancement of Computing in Education (AACE). https ://www.learntechlib.org/primary/p/147396/.

Alderson, C. J. (1980). Native and non-native speaker performance on Cloze tests. *Language Learning, 30*(1), 59-76. https ://doi.org/10.1111/j.1467-1770.1980.tb00151.x

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 115. https ://www.annualreviews.org/doi/pdf/10.1146/annurev.ps.37.020186.000245

Aryadoust, V. (2013). Building a validity argument for a listening test of academic proficiency. Cambridge Scholars Publishing.

Aryadoust, V., Eckes, T., & In'nami Y. (2021) Editorial : Frontiers in language assessment and testing. *Frontiers in psychology, 12,* 1944. https ://doi.org/10.3389/fpsyg.2021.691614

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F. (2005) Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1-34. https ://doi.org/10.1207/s15434311laq0201_1

Blais, J. G., & Laurier M. (1995). The dimensionality of a placement test from several analytical perspectives. *Language Testing, 12*(1), 72-98. https ://doi.org/10.1177/026553229501200105

Bonniol J.-J., & Vial, M. (1997). *Les modèles de l'évaluation*. De Boeck Université.

Brown, J. D. (1997). Computers in language testing : Present research and some future directions. *Language Learning & Technology, 1*(1), 4459. http ://llt.msu.edu/vol1num1/brown/default.html

Canale, M., & Swain, M. (l980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-48. https ://doi.org/10.1093/applin/I.1.1

Carroll, B. J. (1980). *Testing communicative performance*. Pergamon.

Carroll, J. B. (1983) Psychometric theory and language testing. Dans J. W. Jr Oller (dir.), *Issues in language testing research* (p. 80-107). Newbury House.

Carver, R. P. (1974). Two dimensions of tests : Psychometric and edumetric. *American Psychologist, 29*(7), 512–518. https ://doi.org/10.1037/h0036782

Chalhoub-Deville, M. (2010). Technology in standardized language assessments. Dans R. Kaplan (dir.), *The Oxford handbook of applied linguistics* (2e éd., p. 511-526). Oxford University Press. https ://doi.org/10.1093/oxfordhb/9780195384253.013.0035

Chapelle, C. A., & Chung, Y. R. (2010). The promise of NLP and speech processing technologies in language assessment. *Language Testing*, *27*(3), 301-315. https ://doi.org/10.1177/0265532210364405

Chapelle, C. A., & Douglas, D. (2006). *Assessing language through computer technology*. Cambridge University Press.

Chapelle, C. A., & Voss, E. (2016). 20 years of technology and language assessment in Language Learning & Technology. *Language Learning & Technology, 20*(2), 116-128. https://doi.org/10.1177/0265532210364405

Cheng, L., & Curtis, A. (2004). Washback or backwash : A review of the impact of testing on teaching and learning. Dans L. Cheng, Y. L. Wabanabe et A. Curtis (dir.), *Washback in language testing : Research contexts and methods* (p. 3-17). Laurence Earlbaum Associate.

Chomsky, N. (1965). *Aspects of the theory of syntax*. The MIT Press.

Cizek, G. J. (2012). Defining and distinguishing validity : Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*(1), 31–43. https://doi.org/10.1037/a0026975

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments : A practical guide to Kane's framework. *Medical Education, 49*, 560575. https://doi.org//10.1111/medu.12678

Fouly, K. A., Bachman. L. F., & Cziko, G. A. (1990). The divisibility of language competence : A confirmatory approach. *Language Learning, 40*(1), 1-21. https://doi.org/10.1111/j.1467-1770.1990.tb00952.x

Fulcher, G. (2011). Cheating gives life to our test dependence. *The Guardian Weekly*, 14 octobre. http://languagetesting.info/articles/store/cheating.pdf

Germain, C. (1993). *Évolution de l'enseignement des langues : 5 000 ans d'histoire*. CLE international.

Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Language Teaching, 40*(2), 97-118. https://doi.org/10.1017/S0261444807004144

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18*(8), 519–522. https://doi.org/10.1037/h0049294

Hamp-Lyons, L. (1997). Washback, impact and validity : Ethical concerns. *Language Testing*, *14*(3), 295-303. https://doi.org/10.1177/026553229701400306

Hanania, E., & Shikhani, M. (1986). Interrelationships among three tests of language proficiency. *TESOL Quaterly, 20*(1), 97-109. https://doi.org/10.2307/3586391

Hulstijn, J. H. (1985). Second Language proficiency : an interactive approach. Dans K. Hyntenstan et M. Pienemann (dir.), *Modelling and assessing second language acquisition* (p. 373-380). Multilingual Matters.

Hymes, D. (1974*). Foundations in Sociolinguistics : An ethnographic approach*. University of Pennsylvania Press.

International Language Testing Association. (2000). *Code of ethics for ITLA*. http://www.iltaonline.com/code.pdf

International Language Testing Association. (2007). *ILTA guidelines for practice*. https://www.iltaonline.com/page/ILTAGuidelinesforPractice

Isbell, D. R., & Kremmel B. (2020). Test review : Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing, 37*(4), 600-619. https://doi.org/10.1177/0265532220943483

Kane, M. (2006). Validation. Dans R. Brennan (dir.), *Educational measurement* (4e éd., p. 1764). American Council of Education/Praeger.

Kane, M. (2012). Validating score interpretations and uses. *Language Testing, 29*(1), 317. https://doi.org/10.1177/0265532211417210

Klinger, D. A., McDivitt, P. R., Howard, B. B., Munoz, M. A., Rogers, W. T., & Wylie, E. C. (2015). *The classroom assessment standards for PreK-12 teachers.* Kindle Direct Press.

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing, 35*(4), 477-499. https://doi.org/10.1177/0265532217710049

Koh, K. (2014). Authentic assessment, teacher judgment and moderation in a context of high accountability. Dans C. Wyatt-Smith, V., Klenowski et P. Colbert (dir.), *Designing assessment for quality learning* (vol. 1, p. 249-264). Springer. https://doi.org/10.1007/978-94-007-5902-2_16

Lado, R. (1961). *Language Testing : The construction and use of foreign language tests*. Longmans.

Laurier M. (1992) L'application des techniques de tests adaptatifs en français langue seconde. Dans Sauvé L. (dir.), *La technologie éducative d'hier à demain* (p. 109-118). CIPTE/Télé-Université.

Laurier M. (1999). The development of an adaptive test for placement in French. Dans M. Chalhoub-Deville (dir.), *Issues in computer-adaptive testing of reading proficiency* (p. 122-135). Cambridge University Press.

Laurier M., & Baker, B. (2015). The certification of teachers' language competence in Quebec in French and English : Two different perspectives ? *Language Assessment Quarterly*, *12*(1), 10-28. https://doi.org/10.1080/15434303.2014.979349

Laurier M., Lussier, D., & Riel-Salvatore, H. (2021). The development of linguistic profiles and online tests for Quebec health professionals working in English. Dans N. Saville (dir.), *Collated papers for the ALTE 7th International Conference* (p. 171-174). Association of Language Testers in Europe. https://alte.org/resources/Documents/ALTE%207th%20International%20Conference%20Madrid%20June%202021.pdf

Legendre, R. (2000). *Dictionnaire actuel de l'éducation* (2e éd.). Guérin.

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

McNamara, T. (2006). Validity in language testing : The challenge of Sam Messick's legacy. *Language Assessment Quarterly*, *3*(1), 31-51. https://doi.org/10.1207/s15434311laq0301_3

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

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement : Issues and Practice, 16*(2), 16-18.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11), 1012–1027. https://doi.org/10.1037/0003-066X.35.11.1012

Messick, S. (1989). Validity. Dans R. L. Linn (dir.), *Educational measurement* (3e éd., p. 13-104). American Council of Education/Macmillan.

Milliner B., & Barr B. (2020). Computer-assisted language testing and learner behavior. Dans M. Freiermuth et N. Zarrinabadi (dir.), *Technology and the psychology of second language learners and users. New language learning and teaching environments.* Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-34212-8_5

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement : Implications for performance assessment. *Review of Educational Research, 62*(3), 229-258. https://doi.org/10.3102/00346543062003229

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing, 6*(2), 199-215. https://doi.org/10.1177/026553228900600206

Oller, J. W. Jr (1979). *Language tests at school*. Longman.

Papageorgiou, S., & Venessa F. M. (2021) Maintaining access to a large-scale test of academic language proficiency during the pandemic : The launch of TOEFL iBT Home Edition. *Language Assessment Quarterly, 18*(1), 36-41. https://doi.org/10.1080/15434303.2020.1864376

Phakiti, A., & Isaacs, T. (2021). Classroom assessment and validity : Psychometric and edumetric approaches. *European Journal of Applied Linguistics and TEFL, 10*(1), 3-24.

Scallon, G. (2000). *L'évaluation formative.* Éditions du Renouveau pédagogique.

Schissel, J. L., Leung, C., & Chalhoub-Deville, M. (2019). The construct of multilingualism in language testing. *Language Assessment Quarterly*, *16*(4-5), 373-378. https://doi.org/10.1080/15434303.2019.1680679

Scholz, G., Hendricks, D., Spurling, R., Johnson, M., & Vandenberg, L. (1980). Is language ability divisible or unitary ? A factor analysis of 22 English language proficiency tests. Dans J. W. Jr Oller et K. Perkins (dir.), *Research in language testing* (p. 24-33). Newbury House.

Shepard, L. A. (1993). Evaluating test validity. *Review of research in education*, *19*(1), 405-450. https://doi.org/10.3102/0091732X019001405

Shohamy, E. (2001). *The power of tests : A critical perspective on the uses of language tests* (1re éd.). Routledge.

Simon, M. (2011). La qualité d'un instrument d'évaluation de la littératie. Dans M. J. Berger et A. Desrochers (dir.), *L'évaluation de la littératie* (p. 287-314). Les Presses de l'Université d'Ottawa.

Spolsky, B. (1995). *Measured words : The development of objective language testing*. Oxford University Press.

Suvorov, R., & Hegelheimer, V. (2013). Computer-assisted language testing. Dans A. J. Kunnan (dir.), *The companion to language assessment, 2*, 594-613. https://doi.org/10.1002/9781118411360.wbcla083

Tardif, J. (2006). *L'évaluation des compétences : Documenter le parcours de développement*. Chenelière Éducation.

Thomas, N., & Osment, C. (2020). Building on Dewaele's (2018) L1 versus LX dichotomy : The Language-Usage-Identity state model. *Applied Linguistics, 41*(6), 1005-1010. https://doi.org/10.1093/applin/amz010

Viswanathan, U., Lebel, M. E., & Barysevich, A. (2018). Un dispositif pour promouvoir et soutenir l'authenticité des interactions en classe de langue seconde. *Nouvelle Revue Synergies Canada, 11*. https://doi.org/10.21083/nrsc.v0i11.3997

Wall, D., & Alderson, J. C. (1993). Examining washback : the Sri Lankan impact study. *Language testing, 10*(1), 41-69. https://doi.org/10.1177/026553229301000103

Widdowson, H.G. (1978). *Teaching language as communication*. Oxford University Press.

Wiggins, G. (1998). *Educational assessment : Designing assessments to inform and improve student performance*. John Wiley.

Youn, S. J. (2015). Validity argument for assessing L2 pragmatics in interaction using mixed methods. *Language Testing, 32*(2), 199-225. https://doi.org/10.1177/0265532214557113

Zumbo, B. D. (2007). Three generations of DIF analyses : Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly, 4*(2), 223-233. https://doi.org/10.1080/15434300701375832