Mesure et évaluation en éducation

Les standards de performance en éducation

Jean-Guy Blais

Volume 31, Number 2, 2008

URI: https://id.erudit.org/iderudit/1025009ar
DOI: https://doi.org/10.7202/1025009ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

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

Explore this journal

Cite this article

Blais, J.-G. (2008). Les standards de performance en éducation. *Mesure et évaluation en éducation*, *31*(2), 93–105. https://doi.org/10.7202/1025009ar

Article abstract

In education, when one must establish levels that differentiate learning and/or performance of individuals, it is necessary to define standards of performance. There are many proposed methods for developing such standards. These methods share common features, but also have specific characteristics that allow their classification into four main categories, depending on the task requested from the panel of experts. There is no universal method for all situations; the procedures can be improved and, given the increasing frequency of this type of operation, require a high degree of attentive interest.

Les standards de performance en éducation

Jean-Guy Blais
Université de Montréal

Mots clés : Standards, performance, réussite, méthodologie de mesure

En éducation, lorsqu'il faut établir des niveaux pour différencier les apprentissages et les performances des individus, il faut mettre en place des standards de performance. Il existe un grand nombre de méthodes proposées pour élaborer de tels standards. Ces méthodes partagent des points communs, mais ont également des spécificités qui permettent de les classer en quatre grandes catégories en fonction de la tâche demandée aux panélistes experts. Il n'existe pas de méthode universelle pour toutes les situations ; la démarche est perfectible et, étant donné la recrudescence de ce type d'opération, elle exige qu'on s'y intéresse attentivement.

Key words : Standards, performance, success, measurement methodology

In education, when one must establish levels that differentiate learning and/or performance of individuals, it is necessary to define standards of performance. There are many proposed methods for developing such standards. These methods share common features, but also have specific characteristics that allow their classification into four main categories, depending on the task requested from the panel of experts. There is no universal method for all situations ; the procedures can be improved and, given the increasing frequency of this type of operation, require a high degree of attentive interest.

Palavras-chave : Standards, desempenho, sucesso, metodologia de medida

Em educação, quando é necessário estabelecer níveis para diferenciar as aprendizagens e os desempenhos dos indivíduos, é preciso definir standards mínimos de desempenho. Existe um grande número de métodos para elaborar tais standards. Este métodos partilham pontos comuns, mas têm igualmente especificidades que permitem classificá-los em quatro grandes categorias, em função da tarefa solicitada ao painel de peritos. Nâo há um método universal para todas as situações ; o processo é perfectível e, assumindo que este tipo de operação é cada vez mais frequente, reclama um interesse muito atento.

Note de l'auteur – Toute correspondance peut être adressée par courriel à l'adresse suivante : [jean-guy.blais@umontreal.ca].

Introduction

Un observateur attentif de la scène de l'éducation aura remarqué une recrudescence, au cours des dernières années, tant sur la scène nationale que celle internationale, des comparaisons des performances des élèves à des épreuves standardisées. Cet observateur remarquera ainsi une tendance nette à la hausse des opérations de récolte à grande échelle de données sur les apprentissages réalisés par les élèves des ordres primaire et secondaire (Blais, 2004 ; Blais, Laurier & Pelletier, 2001). Ces opérations s'accompagnent souvent d'un processus d'élaboration de « standards » de référence pour observer les mouvements transversaux ou longitudinaux, hausses ou baisses, des résultats des élèves et pour en inférer des conclusions sur l'état du système d'éducation ou même des établissements scolaires[1]. La mise en place de tels standards de performance en éducation n'est pas une opération simple et elle est caractérisée par le fait qu'elle combine le jugement humain avec des modèles psychométriques et des critères de praticabilité (Hambleton & Pitoniak, 2006 p. 435). Une opération de ce type sera donc toujours une combinaison des valeurs sociales d'une époque, de la compréhension du processus d'apprentissage et des avancés techniques de cette même époque (Resnick & Resnick, 1996). Il n'y a pas de méthode unique pour produire des standards, et des méthodes différentes produisent parfois des standards différents (Angoff, 1988). L'interaction continue entre le jugement humain informé, la technique et les données de l'expérience étant au cœur du processus, le vrai test de l'adéquation des standards est leur capacité à être utiles pour la prise de décision en rendant compte le mieux possible de la réalité. Un standard exprime donc certaines valeurs et les méthodes pour le mettre en place peuvent difficilement exprimer une vérité scientifique immuable et à l'épreuve du temps.

Ceci étant dit, la mise en place de standards est une démarche perfectible qui, étant donné la recrudescence de ce type d'opération, exige qu'on s'y intéresse attentivement. Deux aspects reliés aux standards en éducation font ainsi l'objet de la suite de ce texte. D'abord, il est important d'examiner ce que représente le concept de standard de performance en éducation. Ensuite, comme il existe un assez grand nombre de méthodes pour guider l'élaboration de standards, il n'est pas superflu de s'attarder à ce processus et de faire ressortir les points communs et les différences entre des méthodes élaborées et raffinées avec les années pour la production des standards de performance.

Un standard de performance

Dans la langue française, le mot standard sert à qualifier des produits, des situations, des processus normaux dans le sens de ce qui est correct ou adéquat pour un contexte donné (un prix standard, le français standard, une performance standard, *etc.*). Un objet ou une situation standard renvoie à la normalité, à la conformité ou à l'uniformité, c'est-à-dire à quelque chose de courant, d'usuel. Le mot a ainsi pris un sens en français qui l'éloigne de ses origines anglo-saxonnes où il faisait référence aux notions d'excellence et d'exemplarité (Aldrich, 2000).

En éducation, le concept de standard est à l'œuvre principalement lorsqu'il faut établir des niveaux (des échelons gradués) pour différencier les apprentissages ou les performances des individus. Il se réfère, d'une part, au contenu qui fait l'objet de l'apprentissage et, d'autre part, à la performance des élèves dans l'apprentissage de ce contenu. Il existe ainsi un contenu standard ou commun à aborder (*i.e.* le programme), et il y a une exigence minimale à satisfaire pour réussir selon ce qui est attendu socialement. Cette exigence minimale s'appelle un standard de réussite.

Pour plusieurs auteurs (par exemple, Hambleton & Pitoniak, 2006 ; Kane 1994), établir un standard minimal de performance se résume principalement à déterminer un score sur une échelle permettant de classer les candidats dans deux catégories reflétant la compétence et la non-compétence, ou encore dans plusieurs catégories reflétant différents degrés de compétence. À d'autres époques, des expressions comme « scores de césure » (pour *cut-off scores* ou *cut scores*), « niveaux de maîtrise » et « notes de passage » ont également été utilisées pour définir les repères qui établissent la frontière entre les niveaux de compétence. Dans cette optique, le score qui sert à délimiter la compétence minimale constitue une version opérationnelle de ce qui est considéré comme constituant une maîtrise suffisante du contenu. Pour établir un standard de performance minimale, il faut donc trouver une façon de distinguer ce qui est suffisant pour accéder à la « réussite » de ce qui ne l'est pas, tant du point de vue du curriculum ciblé que de la performance lors de l'évaluation des apprentissages. Comme l'idée de standard peut aussi intégrer un continuum avec différents points de référence où un de ces points marque la différence entre ce qui est suffisant et ce qui ne l'est pas, il est aussi possible de créer une échelle avec différents repères ordonnés qui indiqueront que l'apprentissage est marqué, assuré, acceptable, insuffisant ou nettement insuffisant. Dans le même sens, la compétence sera très bien développée, bien développée, assez développée,

peu développée, très peu développée. Dans ce contexte, ces points deviennent autant de sous-standards à définir, des étapes qui indiquent une direction vers ce qui est mieux.

L'élaboration de standards : une multitude d'approches et des points communs

Le grand nombre de méthodes proposées pour élaborer des standards de performance rend difficile l'obtention d'une vue d'ensemble et soulève la question de la classification de ces méthodes dans des ensembles relativement homogènes et cohérents. Déjà Hambleton en 1980 présentait 16 méthodes différentes qu'il classait dans trois catégories : les méthodes reposant d'abord sur le jugement sans accès à des données, les méthodes utilisant des données empiriques comme base de départ et les méthodes combinant ces deux approches (Hambleton, 1980). Plus tard, Berk (1996) relevait 20 méthodes tandis que les différents textes d'un numéro spécial de la revue *Applied Measurement in Education* (1995) permettaient d'en recenser près de 50. Très récemment, Cizek et Bunch (2007) décrivaient 15 méthodes et Hambleton et Pitoniak (2006), pour leur part, faisaient état d'environ 25 méthodes. Il existe ainsi des méthodes pour élaborer des standards qui se basent principalement sur des théories, sur le consensus d'experts ou sur des analyses statistiques. Les standards peuvent être collectifs ou individuels, en ce sens qu'ils visent soit l'ensemble des personnes engagées dans un cycle d'études ou un programme, soit une épreuve précise qui doit être réussie par des individus pour accéder à un autre niveau. Lorsque la démarche d'élaboration d'un standard vise une épreuve en particulier, une distinction peut être faite, parmi les méthodes, entre celles qui centrent les opérations sur les tâches à réaliser par les candidats et celles qui se basent sur des travaux déjà réalisés par ceux-ci (*test-centered method* et *examinee-centered method*).

Longtemps, les méthodes pour établir des standards ont été divisées selon les deux catégories mentionnées ci-dessus, *i.e.* celles qui étaient centrées sur l'individu et celles centrées sur le test. Cependant, cette classification est devenue limitée avec les années, par l'entrée en scène de nouvelles méthodes, et demandait une révision. Ainsi, à la suggestion de Hambleton, Jaeger, Plake et Mills (2000), les méthodes pour établir des standards de performance pourraient maintenant être classées en quatre grandes catégories reposant sur le travail exigé des experts panélistes (les grands principes des méthodes de chacune de ces catégories sont présentés ci-dessous) :

- les méthodes qui impliquent l'examen des items et des rubriques d'attribution des scores,
- les méthodes qui impliquent l'examen des candidats,
- les méthodes qui impliquent l'examen du travail effectué par les candidats,
- les méthodes qui impliquent l'examen par les experts de profils de scores.

D'après le document *Standards for educational and psychological testing* (AERA, APA, NCME, 1999, p. 53), il n'existe pas de méthode universelle pour tous les tests et toutes les situations, comme il n'existe pas de démarche unique pour démonter la validité des standards proposés avec une méthode particulière. Cependant, il est important de retenir que peu importe le choix de la méthode celle-ci partage probablement plusieurs points communs avec d'autres démarches méthodologiques ayant les mêmes objectifs (Cizek & Bunch, 2007, p. 35-41 ; Hambleton, 1998 ; Hambleton & Pitoniak, 2006, p. 436-439). La suite de cette section permet donc, dans un premier temps, de prendre connaissance des étapes communes aux différentes méthodes et, ensuite, de mieux saisir les caractéristiques des méthodes de chacune quatre catégories présentées ci-dessus. Le tableau 1, adapté de Hambleton et Pitoniak (2006), offre un résumé des principales caractéristiques des méthodes de ces catégories.

Les points communs aux différentes méthodes

Préalablement à toute opération d'élaboration de standards de performance, il faut préciser l'objectif de l'opération. Cette étape est déterminante car elle oriente par la suite les décisions concernant les caractéristiques désirées de l'instrument de mesure, les conditions d'administration, le type de données à récolter et le jugement sur les critères de validité de la procédure. Ensuite, il faut choisir une méthode, et ce, en fonction des connaissances ou des compétences visées, du type d'items privilégiés (réponse choisie ou construite, par exemple), du temps et des ressources disponibles, de l'expérience des responsables et des résultats de recherches sur cette méthode. Après le choix de la méthode, il faut déterminer le nombre de niveaux de performance à établir (2, 3, 4, *etc.*) et le nombre d'items ou de tâches dans les épreuves utilisées. Plus le nombre de catégories est élevé et plus la tâche devient difficile pour les « experts », peu importe la méthode. Moins il y a d'items ou de tâches dans une épreuve et plus il est difficile de proposer plusieurs points de coupure correspondant à autant de degrés de performance[2]. Les étiquettes et les descriptions associées aux différents niveaux de performance font aussi l'objet d'une attention particulière dans toutes les méthodes. Ces repères informent succinctement du résultat obtenu et peuvent être sujets à différentes

interprétations s'ils sont mal définis. Dans la plupart des méthodes, il faut donc définir en mots, habituellement un ou deux courts paragraphes, ce que constitue la performance attendue pour chacun des niveaux établis. Ces descriptions portent également le nom de « rubriques » et permettent de documenter le jugement en offrant un résumé de ce que le candidat devrait faire pour être classé dans l'une ou l'autre des catégories (Blais, Laurier & Rousseau, sous presse). Elles constituent autant de performances standards auxquels la performance observée est comparée.

Il faut également sélectionner et entraîner les membres du panel, les « experts », qui seront appelés à porter des jugements sur les items ou les candidats. Selon la méthode, le travail des membres du panel sera plus ou moins intensif et pourra s'étaler sur plusieurs jours. Mise à part leur disponibilité pour une certaine période, un des premiers critères pour le recrutement des membres du panel et pour décider de leur nombre est leur « représentativité » de la population des personnes considérées compétentes pour effectuer la tâche. Les définitions que l'on adopte de la représentativité et de la compétence peuvent varier selon le contexte et les objectifs, mais le critère ultime de l'adéquation du panel pour la tâche, peu importe la méthode, est le fait que les standards ne devraient pas varier si un autre groupe de panélistes aussi représentatif et aussi compétent répétait le processus (AERA, APA, NCME, 1999, p. 54). Pour certaines tâches, il est crucial que les panélistes aient une expérience plus que raisonnable des contenus et des élèves qui y sont exposés. Parce qu'ils sont bien placés pour produire cette expertise, les enseignants sont ainsi souvent sollicités pour agir comme juges ou experts dans différentes occasions où il faut statuer sur ce que savent et peuvent faire les élèves. Après la sélection des membres du panel, une période de formation doit être planifiée pour induire chez tous une vision homogène de la tâche à exécuter. Il s'agit de faire en sorte que les personnes retenues comprennent les objectifs de l'opération, les concepts importants, le rôle des rubriques et le type de jugement qu'elles auront à porter. Très souvent, des prototypes d'items et de productions (par exemple, des textes écrits) sont utilisés pour aider les membres du panel et servir de référence lors de la formation pour déterminer les points d'ancrage de l'échelle de compétence. Dans le même sens, certaines méthodes prônent la communication aux membres du panel de données empiriques pour les aider à réviser leurs décisions. Finalement, une dernière étape commune à toutes les méthodes consiste en la compilation des cotes en utilisant généralement la moyenne ou la médiane des cotes attribuées par les panelistes.

Des méthodes qui impliquent l'examen des items ou des tâches

La plupart des méthodes de cette catégorie s'inspirent de celle d'Angoff (1971). Elles s'appliquent mieux lorsque les items sont à réponse choisie (QCM par exemple) et elles sont parmi les méthodes les plus étudiées à ce jour. De façon générale, la procédure demande aux panélistes d'estimer la probabilité qu'un candidat se situant à la limite ou au seuil de la réussite (*borderline candidate*) puisse réussir une tâche donnée. On y retrouve plusieurs variantes de la méthode d'Angoff, de même que les méthodes de Ebel (1972), Jaeger (1995) et Nedelksy (1954), les méthodes de projection et celles de consensus. Dans la méthode d'Angoff et ses multiples variantes, les membres du panel doivent examiner attentivement chacun des items et produire une estimation de la probabilité qu'un candidat avec la compétence minimale, un candidat se situant au seuil de réussite, puisse répondre correctement à un item donné. Les estimations sont ensuite compilées de différentes façons pour produire le standard de référence du panel. Dans plusieurs variantes de la méthode, le processus est itératif et les membres du panel sont invités à revoir leurs estimations à la lumière de renseignements supplémentaires tels que des données empiriques sur les items (par exemple la valeur de la difficulté d'un item pour la population visée) ou encore à la suite des échanges avec les autres panélistes. Les méthodes de cette catégorie ont été souvent critiquées parce que la tâche demandée aux panélistes serait trop difficile à accomplir avec une précision minimale et constituerait une tâche cognitive non raisonnable (Sheppard, Glaser, Linn & Bohrnstedt, 1993). Il semble que le degré de précision avec lequel les panélistes peuvent estimer la probabilité qu'un candidat obtienne la bonne réponse à un item soit en lien avec les caractéristiques des panélistes, avec le type de données empiriques fournies et avec le niveau de difficulté réel des items considérés (voir Brandon, 2004 et Stone, 2004). Différentes variantes à la méthode d'Angoff proposées ces dernières années visent la production de plus d'un point de références, *i.e.* des standards sur une échelle en catégories ordonnées (*i.e.* une échelle polytomique), afin de mieux tenir compte de l'existence de plusieurs paliers de performance. Ces dernières années, la méthode du « signet » (*bookmark method*) et la méthode du « consensus direct » semblent avoir la faveur de plusieurs spécialistes dans le domaine (voir Cizek & Bunch 2007, chap. 10).

Des méthodes qui impliquent un jugement direct des candidats

Les méthodes de cette catégorie exigent des panélistes qu'ils produisent un jugement directement sur les candidats et qu'ils les classent en deux ou plusieurs catégories de compétence (par exemple : réussite, frontière, échec). Généralement, lorsque les panélistes amorcent leur travail, ils n'ont aucune information sur la performance préalable des candidats à une épreuve quelconque et doivent faire leur travail à partir de leur seule connaissance du candidat. Mais, par la suite, les standards sont établis en combinant le jugement direct sur le candidat et le résultat du candidat à une épreuve particulière en lien avec le contenu visé par l'opération. La principale difficulté relevée quant à l'application des méthodes de cette catégorie réside dans le fait qu'elles nécessitent le recrutement de panélistes qui connaissent suffisamment les candidats. Dans un contexte scolaire, des enseignants peuvent probablement réaliser la tâche adéquatement, mais les résultats sont difficilement généralisables. Cependant, il s'agit d'une tâche qui leur est familière, donc moins difficile pour eux que pour d'autres, et le temps de formation est moins long.

Des méthodes qui impliquent l'examen du travail effectué par les candidats

Alors que les méthodes des catégories précédentes sont plus performantes lorsque les tâches sont constituées d'items à réponse choisie, les méthodes de cette troisième catégorie sont plus appropriées lorsque les items sont à réponse construite. Les méthodes sont très semblables et se distinguent en fonction du type d'items dans une épreuve (uniquement à réponse construite ou une combinaison avec des items à réponse choisie), de la correction (globalement ou item par item, donc avec une pondération possible selon les items), de la nature de la tâche pour les panélistes (par exemple, distinguer les meilleures réponses, les réponses *borderline* et les moins bonnes réponses) et selon la méthode pour calculer le standard de performance final. La plupart des méthodes adoptent une perspective holistique plutôt qu'analytique. Le jugement est donc surtout global et ne s'attarde pas à documenter la performance pour des items/tâches/critères en particulier. Certaines méthodes proposent toutefois une procédure hybride où plusieurs jugements holistiques sont faits sur les principales composantes du test pour être ensuite combinés en un jugement global (voir Hambleton & Pitoniak, 2006, p. 447). De façon générale, ces méthodes apparaissent plus simples aux panélistes car elles leur demandent une tâche relativement facile à exécuter consistant à ordonner les réponses des candidats sur un continuum de performance (une tâche que les enseignants, par exemple, ont l'habitude d'accomplir).

Des méthodes qui impliquent l'examen par les experts de profils de scores

Dans cette dernière catégorie se retrouvent des méthodes développées plus récemment et visant des performances complexes et multidimensionnelles où il faut tenir compte du profil de scores d'un candidat. Les panélistes doivent ainsi examiner des profils théoriques de scores et déterminer une catégorie de performance à laquelle chaque profil devrait être rattaché. Par exemple, dans une situation ou il y a quatre critères cotés de 1 à 4, un profil réussite pourrait être 4, 3, 3, 4, et un profil échec pourrait être 1, 1, 1, 2. Ces profils sont, en quelque sorte, exemplaires et servent de guides pour la classification des candidats selon les standards de performance. Dans certains cas, les panélistes doivent aussi déterminer les profils de scores à la frontière du succès ou de l'échec (*borderline scores profiles*). Dans l'exemple précédent, cela pourrait être un profil 2, 3, 2, 2, par exemple. Deux méthodes sont citées plus fréquemment dans cette catégorie : il s'agit de la méthode du « profil dominant » et de la méthode de la « capture du jugement » (*judgmental policy capturing method*). Pour cette dernière méthode, l'objectif est de découvrir la position implicite de chacun des membres du panel quant aux compétences et habiletés exigées pour réussir. Les méthodes de cette catégorie ont surtout été utilisées dans des contextes de certification professionnelle (avec des enseignants par exemple ; voir Jaeger, 1995).

Au-delà de ces quatre catégories de méthodes, il est possible de recenser d'autres propositions pour des méthodes hybrides dites de « compromis » (les méthodes de Hofstee et de Beuk par exemple) et des méthodes mixtes, quantitatives et qualitatives, pour l'élaboration de rubriques (Blais, Laurier & Rousseau, sous presse). Enfin, la question de la compensation revient sur le tapis pour toutes les méthodes. Doit-on permettre qu'un résultat faible soit compensé par un autre plus élevé en tenant compte uniquement de la somme des scores (ou cotes) ou doit-on exiger des candidats qu'ils atteignent un standard minimal pour chacune des parties ou tâches d'une épreuve ? Par exemple, avec quatre critères pour juger de la performance, un profil de 1, 1, 4, 4, donne une somme de 10 sur 16 qui placerait le candidat dans la catégorie de la réussite même si, de toute évidence, il montre des signes de difficulté pour ce que représentent les critères 1 et 2. Finalement, les préoccupations au sujet des candidats ayant des handicaps ont mené à la mise en place de procédures « alternatives » pour l'évaluation des apprentissages et même si, théoriquement, toutes les méthodes pour élaborer des standards pourraient être appliquées pour les individus de cette population, il va de soi que cela dépend de la procédure utilisée pour récolter les données.

Tableau 1
Récapitulatif des caractéristiques des méthodes pour l'élaboration de standards selon les quatre grandes catégories de Hambleton et Pitoniak (2006)

1. Des méthodes qui impliquent l'examen des items ou des tâches	
Angoff, Ebel, Nedelsky, Jaeger	Les panélistes estiment la probabilité que le candidat se situant au seuil de la réussite réponde correctement à chacun des items (items à réponse choisie, items polytomiques, etc.).
Signet	Les panélistes examinent un ensemble d'items placés en ordre de difficulté et placent un signet pour séparer les items en fonction de la probabilité de réussite d'un candidat se situant au seuil de la réussite.
Consensus direct	Les panélistes travaillent avec des regroupements d'items et proposent une estimation du nombre d'items que le candidat au seuil de la réussite sera capable de réussir dans chacun des regroupements.
2. Des méthodes qui impliquent un jugement direct des candidats	
Groupes contrastés et groupes frontières	Les panélistes examinent les résultats des candidats et identifient un groupe dont les membres sont nettement au-dessus d'un standard de performance donné et un groupe dont les membres sont nettement en-dessous du standard.
3. Des méthodes qui impliquent l'examen du travail effectué par les candidats	
Examen item par item	Pour chaque item, les panélistes examinent un échantillon des productions des candidats et sélectionnent les productions qu'ils associent à un candidat au seuil de la réussite.
Examen holistique	À partir de l'ensemble de la production d'un candidat, les panélistes placent les candidats dans différentes catégories de performance.
4. Des méthodes qui impliquent l'examen par les experts de profils de scores	
Capture du jugement	Les panélistes examinent des profils de scores et associent les différents profils à des niveaux de compétence. Les données permettent ensuite d'établir un portrait du jugement de chacun des panélistes.
Profil dominant	Les panélistes examinent des profils de scores et tentent d'en arriver à un consensus sur la politique à retenir pour déterminer le standard.
Regroupement d'items	Les panélistes examinent les schémas de réponses à des items à réponse choisie et placent ces schémas sur une échelle ordonnée de quatre points.

Conclusion

Établir des standards en éducation est une opération qui vise essentiellement à étalonner des valeurs subjectives, une opération qui a plus à voir avec les laboratoires de psychophysique du XIX^e^ siècle qu'avec la mesure de la distance par exemple. À la différence que pour la mise en place de standards de performance en éducation il n'y pas de référent externe objectif et stable, comme un mètre, auquel pourraient se référer ceux et celles qui sont appelés comme juges ou experts pour échelonner les apprentissages et les compétences selon des standards de performance. L'élaboration, le pilotage ou l'évaluation d'un standard dépendent ainsi fortement des méthodes et des techniques d'analyse mises à contribution. Dans les écrits sur le sujet, il est largement reconnu qu'il y a un effet dû à la méthode et la conclusion qui revient continuellement est que des méthodes différentes donnent des résultats différents qui peuvent même être en contradiction (Stone, 2004). L'opération n'est donc pas gratuite, nous n'avons pas de mètre et pas de balance, tout doit être construit sur mesure à partir de notre expérience et de celle des autres. D'un côté, cela peut sembler inquiétant car il y a plusieurs avenues possibles, mais d'un autre côté, cette situation est aussi très stimulante car il y a beaucoup à faire dans une période ou il y a une forte demande en la matière.

Le grand nombre de propositions pour le développement de standards de performance en éducation illustre bien, d'une part, le foisonnement d'idées sur le sujet et, d'autre part, la diversité des situations où il y a une demande pour la production de standards de référence pour la compétence. Malgré ce grand nombre de propositions, plusieurs problèmes intéressants et importants sont toujours à la recherche de solutions adéquates et pratiques. Parmi ces problèmes, il faut mentionner l'appariement des standards pour un même contenu selon différents niveaux scolaires (mathématiques en 4^e^, 6^e^ et 8^e^ années par exemple) et pour des contenus différents pour le même niveau scolaire (histoire et mathématiques en 5^e^ année par exemple). En effet, pour effectuer un suivi cohérent et bien piloter un système d'éducation, il faut s'assurer que les standards peuvent non seulement servir à décrire l'état des lieux mais aussi servir à comparer différentes performances. Un autre sujet d'étude important à mentionner est celui de l'influence des caractéristiques des panélistes sur les standards élaborés. Il faut évidemment que les panélistes soient représentatifs et possèdent les compétences pour réaliser la tâche demandée. Mais au-delà de ces exigences, il faut se demander quelles sont caractéristiques idiosyncratiques des panélistes pouvant influencer indûment le processus, car il s'agit

après tout de jugements humains et de valeurs humaines. Finalement, l'appareillage technique d'une méthode ne dispense pas de la nécessité de bien présenter et de bien communiquer les résultats de l'opération aux décideurs et aux responsables, afin que ceux-ci soient en mesure de prendre des décisions éclairées et valides.

NOTES

1. L'éducation n'est pas le seul secteur de l'activité humaine où on développe des standards. Le secteur manufacturier y est soumis depuis longtemps (le contrôle de la qualité) et les secteurs de l'environnement et de la santé ont leurs lots de standards de dangerosité.
2. À l'ordre primaire, il est évident que la longueur des épreuves est limitée par le contexte et la capacité des élèves à bien performer sur une longue durée.

RÉFÉRENCES

Aldrich, R. (2000). Educational standards in historical perspective. In H. Goldstein & A. Heath (dir.), *Educational Standards* (pp. 39-67). Londres : The British Academy.

American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME) (1999). *Standards for educational and psychological testing*. Washington, D.C. : American Psychological Association.

Angoff, W.H. (1971). Scales, norms and equivalent scores. In R.L. Thorndike (dir.), *Educational Measurement* (2e éd., pp. 508-600). Washington, DC : ACE.

Angoff, W.H. (1988). Proposals for theoretical and applied development in measurement. *Applied Measurement in Education*, *1*(3)*,* 215-222.

Applied Measurement in Education (1995). Numéro spécial, *8*(1).

Berk, R.A. (1996). Standard setting : the next generation*. Applied Measurement in Education*, *9*(3), 215-236.

Blais, J.-G. (2004). L'obligation de résultats à la lumière de l'interaction entre le quantitatif et le social. In C. Lessard & P. Meirieu (dir.), *L'obligation de résultats en éducation* (pp. 123-144). Québec : Les Presses de l'Université Laval.

Blais, J.-G., Laurier, M., & Pelletier, G. (2001). Les indicateurs comme outil de régulation en éducation : entre culture de chercheur et culture de décideur. In G. Pelletier (dir.), *Autonomie et décentralisation en éducation : entre projet et évaluation* (pp. 131-148). Montréal : Édition de l'AFIDES.

Blais, J.-G., Laurier, M., & Rousseau, C. (sous presse). Deriving Proficiency Scales from Performance Indicators Using the Rasch Model. In E. Smith (dir.), *Advances in standard setting*. Maple Grove, MN : JAM Press.

Brandon, P.R. (2004). Conclusions about frequently studied modified Angoff standard-setting topics. *Applied Measurement in Education*, *17*, 59-88.

Cizek, G.J. (dir.) (2001). *Setting performance standards : Concepts, methods and perspectives*. Mahwah, NJ : Lawrence Erlbaum.

Cizek, G.J., & Bunch, M.B. (2007). *Standard setting : a guide to establishing and evaluating performance standard on tests*. Thousand Oaks, CA : Sage.

Ebel, R.L. (1972). *Essentials of educational measurement*. Englewood Cliffs, NJ : Prentice-Hall.

Hambleton, R.K. (1980). Test score validity and standard setting methods. In R.A. Berk (dir.), *A guide to criterion-referenced test construction* (pp. 80-128). Baltimore, MD : Johns Hopkins University Press.

Hambleton, R.K. (1998). Setting performance standards on achievement tests : Meeting the requirement of title 1. In L. Hansche (éd.), *Handbook for the development of performance standards : Meeting the requirements of Title 1* (pp. 87-114). Washington, DC : Council of Chief State School Officers.

Hambleton, R.K., Jaeger, R.M., Plake, B.S. & Mills, C.N. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement*, *24*, 355-366.

Hambleton, R.K., & Pitoniak, M.J. (2006). Setting performance standards. In R.L. Brennan (dir.), *Educational Measurement* (4e éd., pp. 413-470). Washington, DC : ACE/Praeger.

Jaeger, R.M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education*, *8*, 15-40.

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research*, *64*(3), 425-461.

Kane, M.B., & Mitchell, R. (dir.) (1996). *Implementing performance assessment : Promises, problems and challenges*. Mahwah, NJ : Lawrence Erlbaum.

Nedelsky, L. (1954). Absolute grading for objective tests. *Educational and Psychological Measurement*, *14*, 3-19.

Resnick, D.P., & Resnick, L.B. (1996). Performance measurement and the multiple functions of educational measurement. In M.B. Kane & R. Mitchell (dir.), *Implementing performance assessment : Promises, problems and challenges* (pp. 23-38). Mahwah, NJ : Lawrence Erlbaum.

Sheppard, L.A., Glaser, R., Linn, R. & Bohrnstedt, G. (1993). *Setting performance standards for student achievement*. Standford, CA : National Academy of Education.

Stone, G.E. (2004). Objective standard setting. In E.V. Smith & R.M. Smith (dir.), *Introduction to Rasch measurement* (pp. 445-459). Maple Grove, MN : JAM Press.