Validité globale d'une stratégie de testing adaptatif de maîtrise pour fins de certification scolaire au Québec[*]

Réjean Auger
Serge P. Séguin
université du québec à montréal

L'étude concerne l'évaluation de la validité globale d'une stratégie de testing adaptatif informatisé faisant appel au modèle dichotomique logistique à trois paramètres de la théorie des réponses aux items. À tour de rôle, les différentes facettes de la validité de la stratégie de testing sont appréciées selon une grille d'analyse comprenant deux axes principaux: les assises (évidences empiriques et conséquences en regard des décisions) et les fonctions (interprétation et utilisation). Suite aux diverses analyses, les auteurs croient que le testing adaptatif informatisé dans un contexte de certification scolaire au Québec est bel et bien possible selon un devis spécifique de réalisation. En se basant sur les améliorations à apporter en regard de la validité globale, les auteurs proposent une prospective d'activités de recherche et de développement notamment en fonction du modèle de Rasch (1980), de stratégies différentes de testing adaptatif et d'applications dans d'autres champs disciplinaires comme la psychologie.

This study deals with the evaluation of the comprehensive validity of a computerized adaptive testing strategy using a dichotomous three-parameter logistic model from item response theory. The different facets of the testing strategy's validity are assessed in turn on an analytical grid with two principal axes: basis (evidential and consequential) and function (interpretation and use). Following various analyses, we believe that specifically designed computerized adaptive testing is certainly possible in a context of academic certification in Quebec. Basing our proposals on improvements to be made to the comprehensive validity, we suggest future research and development using Rasch's (1980) model, as well as different adaptive testing strategies and applications in other disciplines like psychology.

STRATÉGIE DE TESTING ADAPTATIF INFORMATISÉ

Les pratiques actuelles d'évaluation des apprentissages font ressortir un mode dominant de prise d'information. Il s'agit, habituellement, de tests papier-crayon

[*]Ces recherches ont été subventionnées par le Fonds pour la Formation de Chercheurs et l'Aide à la Recherche (FCAR #94-NC-0963) et le Conseil de Recherches en Sciences Humaines du Canada (CRSH #410-93-0609).

de formats traditionnels administrés à tous les sujets du groupe visé et analysés selon la théorie classique des scores: la précision de la mesure de chaque individu est alors tributaire du groupe et d'un coefficient global de fidélité. Une façon de contrer cette limite est l'emploi de la théorie des réponses aux items (TRI) qui permet d'estimer plus adéquatement la précision de la mesure individuelle. Compte tenu des connaissances actuelles (Wainer et al., 1990; Weiss et Vale, 1987), seul le testing adaptatif sur ordinateur semble approprié pour obtenir la même précision de l'estimé d'habileté individuelle pour tous les examinés. Dans un contexte de certification scolaire, un testing adaptatif ne peut s'effectuer sans l'apport de la TRI. En effet la TRI propose que tout item soit caractérisé en termes d'une probabilité de réussir l'item en fonction de l'habileté du sujet qui y répond: c'est donc la relation entre l'habileté individuelle et la probabilité de réussir l'item qui est au centre de la théorie des réponses aux items. Plusieurs modèles logistiques sont disponibles pour paramétriser les items allant des modèles dichotomiques de un à trois paramètres (difficulté, discrimination, pseudo-hasard) aux modèles polytomiques (catégoriels, gradués de type Likert, etc.). Le modèle choisi pour la paramétrisation des items est aussi utilisé pour l'estimation des habiletés individuelles, permettant que ces dernières et les paramètres de difficulté des items soient exprimés sur une même échelle.

Le testing adaptatif désigne:

> une stratégie de sélection d'items in situ qui consiste à n'administrer à chaque individu que les items permettant de bien mesurer son niveau d'habileté: toute réussite est suivie par un item plus difficile et tout échec, par un item plus facile parmi les items non encore répondus; la sélection d'items est opérée jusqu'à ce que le niveau d'habileté de l'individu soit estimé avec un maximum de précision. (Auger et Séguin, 1992, p. 106)

Des études théoriques, de simulation et empiriques (Hambleton, Swaminathan et Rogers, 1991; Wainer et al., 1990; Weiss et Vale, 1987) confirment qu'une telle stratégie diminue la durée du testing et le nombre d'items administrés, ou augmente la précision de la mesure.

Auger et Séguin (1992) ont déjà fait rapport d'une étude d'une stratégie de testing adaptatif informatisé faisant appel au modèle dichotomique logistique à trois paramètres (Birnbaum, 1968). Dans cette étude, l'algorithme utilisé pour l'estimation du niveau d'habileté est une procédure d'estimation bayesienne. La procédure bayesienne produit un estimé a posteriori qui prend en compte un estimé a priori et l'information du dernier item administré. Comme l'estimé a posteriori est fonction de l'a priori, on doit assurer un point d'entrée valide. C'est pourquoi un prétest (maximum de six items) est d'abord utilisé pour déterminer ce point d'entrée. Ce prétest utilise la note d'école de l'élève comme point de départ, celle-ci étant répartie selon six grandes catégories. Comme on cherche avant tout la précision de l'estimation du niveau individuel d'habileté, le critère d'arrêt du testing en est un de précision de l'estimé d'habileté et le nombre

d'items administrés diffère d'un individu à un autre. Suite à l'arrêt du testing, une procédure automatisée reliée au processus décisionnel est enclenchée. À ce moment, l'estimé d'habileté de l'examiné est comparé aux zones de décision préalablement établies par une méthode d'établissement de scores de césure (Angoff, 1984) et d'animation en convergence intégrée à la méthode d'Angoff (Auger, 1989), transposées sur l'échelle d'habileté. Les règles de décision s'énoncent comme suit: on déclare un individu en situation de réussir l'ensemble concerné des apprentissages si son estimé d'habileté est plus grand ou égal à la limite inférieure de la zone d'incertitude du score de césure; on l'estime en situation d'échouer l'ensemble concerné des apprentissages si son estimé d'habileté est plus petit que la valeur de la limite inférieure de la zone d'incertitude du score de césure.

OBJECTIF

La stratégie de testing adaptatif informatisé proposée par Auger et Séguin (1992), utilisant une interprétation critérielle, est dite de maîtrise, d'où son acronyme TAM. Expérimentée dans un contexte d'évaluation sommative des apprentissages à des fins de sanction des études au Québec, elle a pour buts: (1) d'analyser les qualités métrologiques des items et des résultats obtenus; (2) de vérifier l'efficience statistique des résultats du TAM, comparativement aux résultats du testing conventionnel (un seul examen de type papier-crayon administré à tous), en termes de décisions de maîtrise ou non des apprentissages; et (3) d'étudier les conditions de praticabilité du TAM au Québec. Les lecteurs intéressés trouveront dans Auger et Séguin (1992) les informations pertinentes à la problématique de départ, au devis méthodologique, aux résultats détaillés ainsi qu'à la conclusion sur la praticabilité du TAM appliqué aux contraintes québécoises.

L'objectif ici recherché est celui de ré-évaluer les résultats présentés par Auger et Séguin (1992) en regard d'un nouveau cadre paradigmatique d'analyse, dit de validité globale, dû à Messick (1988). Ce cadre d'analyse permet d'apprécier la validité de la stratégie du TAM dans un contexte plus global à l'intérieur d'une démarche évaluative.

VALIDITÉ GLOBALE

Les pratiques évaluatives sont tributaires de multiples contraintes rencontrées dans le milieu scolaire (Auger et Dassa, 1992), contraintes qui influent sur la validité des tests. Bien qu'interreliées, ces contraintes peuvent se regrouper en deux catégories selon qu'elles relèvent de facteurs internes ou externes au processus d'évaluation.

Les facteurs internes concernent, entre autres, les intentions et les décisions d'ordre pédagogique et d'ordre administratif prises par les enseignants en ce qui

a trait aux apprentissages visés. Du fait que ces intentions portent sur les objectifs, les contenus et les programmes, les facteurs internes trouvent leur origine dans les objets d'évaluation issus des objets d'apprentissage.

Quant aux facteurs externes, ils portent surtout sur l'effet des décisions des enseignants et des administrateurs scolaires au regard de la compétence scolaire ainsi que sur les qualités prédictives des tests et des examens. Qui plus est, l'ensemble de ces décisions doit assurer un traitement équitable pour tous, c'est-à-dire que ces décisions ne doivent entraîner aucun biais envers des sous-populations.

Le contexte d'une démarche évaluative réfère au but visé par l'évaluation, la mise en place d'une instrumentation pour la cueillette d'informations, l'interprétation des scores au test, l'appréciation de l'écart entre l'attendu et l'observé, de même que la prise de décision. L'ensemble de ces étapes est souvent résumé dans le trio suivant: information, jugement, décision. À partir de ce processus d'évaluation, on peut affirmer que le concept traditionnel de la validité ne concerne qu'une partie des préoccupations inhérentes à la démarche évaluative. La validité, au cours des années, a pris plusieurs formes dépendamment des préoccupations du jour. On en retrouve 38 définitions spécifiques répertoriées par Legendre (1993). Cette spécificité ressemble plus à un éclatement et indique, en quelque sorte, que le concept de validité recouvre une réalité multiple.

Qui plus est, aucun des trois types généraux de validité proposés par les "Standards for Educational and Psychological Tests" (American Psychological Association, 1985), soit la validité de construit, la validité reliée à un critère et la validité de contenu, ne reflète la nature globale de l'évaluation des pratiques en milieu scolaire.

Il faut donc une définition de la validité qui prenne en compte l'ensemble des composantes de la démarche d'évaluation, agissant en quelque sorte comme un unificateur du processus d'évaluation pour garantir la qualité des informations, des jugements et des décisions. Cela implique de définir la validité globale en termes d'arguments de validation.

Ainsi, dans le contexte de débats et de mises en commun de textes de 18 spécialistes sur la validité, initiés et édités par Wainer et Braun (1988), Messick (1988) définit la validité d'un test "as an overall evaluative judgment of the adequacy and appropriateness of inferences and actions based on test score" (p. 42) et conclut qu'elle doit prendre en compte quatre aspects: la pertinence et l'utilité du test, l'interprétation des scores du test, l'importance accordée aux actions fondées sur les scores du test ainsi que les conséquences sociales de l'utilisation de ces scores. Épousant l'idée émise par Messick, la validité n'est plus une caractéristique intrinsèque d'un instrument de mesure, mais elle est elle-même évaluation ou jugement. Dans cette ligne de pensée et prolongeant Auger et Dassa (1992), nous proposons la définition suivante: *La validité globale est une évaluation de la contribution de l'ensemble des validités spécifiques et*

des protocoles mis en place, fondée sur des évidences empiriques et sur un rationnel théorique, à l'adéquation et à la justesse des inférences et des actions d'informations recueillies et des scores au test.

Selon cette conception de la validité globale, la pertinence d'un test, l'interprétation des résultats et l'utilité des scores au test sont indissociables et doivent être intégrées au contexte d'une démarche évaluative. Cette conception de la validité globale d'un test dans une démarche évaluative permet d'évaluer chacune des étapes de cette démarche en regard des assises (évidences empiriques et conséquences des décisions) et des fonctions (interprétation et utilisation). Les assises sont constituées des fondements théoriques sur lesquels repose toute mesure (test) d'un objet d'étude ainsi que des conséquences des décisions prises sur la base des informations ou des scores issus du test. Les fonctions s'adressent à l'interprétation des informations ou des résultats au test, ainsi qu'à la pertinence d'utiliser le test en regard du contexte et des objectifs de la démarche évaluative.

Le jugement ou l'évaluation de validité globale d'un test résulte alors d'un processus de dynamique rétroactive entre les preuves et les conséquences relatives à l'interprétation et à l'utilisation des tests. Cette dynamique impose d'intégrer l'analyse de chacune des étapes de la démarche évaluative. Une grille d'analyse permettant de croiser les fonctions et les assises de la validité, telle celle proposée à la figure 1, facilite la visualisation de l'ensemble des préoccupations de la validité globale. Cette grille, lorsque confrontée à une typologie générale de la validité des tests, permet une critique des pratiques usuelles ainsi que l'identification de pratiques novatrices de testing pouvant contribuer à améliorer la validité globale. C'est à partir de cette grille que, dans la suite de cet article, la validité globale du TAM est évaluée.

	Fonctions	
Assises	*Interprétation (validité interne)*	*Utilisation (validité externe)*
Évidences empiriques	A	C
Conséquences des décisions	B	D

FIGURE 1

Les composantes de la validité globale

	Fonctions	
Assises	*Interprétation (validité interne)*	*Utilisation (validité externe)*
Évidences empiriques	(A) • Validité de contenu de la banque d'items (A1) • Adéquation du modèle de paramétrisation des items de la banque (A2) • Transformation des paramètres des items sur une même métrique selon la TRI (A3) • Vérification de l'efficience statistique du testing adaptatif selon une fonction d'information (A4) • Validité discriminante (A5)	(C) • Validité corrélationnelle reliée à des critères (C1) • Degré de concordance des décisions entre le testing adaptatif et le testing conventionnel (C2) • Écologique: respect des contraintes du milieu dans l'expérimentation et respect de la représentativité du programme d'études (C3) • Fonctionnement différencié entre les garçons et les filles, les initiés et les non-initiés à l'utilisation de micro-ordinateurs (C4)
Conséquences des décisions	(B) *Protocoles méthodologiques* • Règles de décision (B1) • Établissement d'un critère de passage ou d'un score de césure (B2) • Choix des juges et application rigoureuse du protocole (B3) • Évaluation globale du profil attendu de compétence scolaire (B4)	(D) *Pertinence et cohérence méthodologiques* • Pertinence et cohérence de la méthodologie utilisée en regard des objectifs de la recherche et de son contexte spécifique (D1) • Pertinence et cohérence des décisions concernant les populations scolaires en regard des finalités du testing (D2)

FIGURE 2

Les diverses facettes de la validité globale appliquée au TAM

LA VALIDITÉ GLOBALE APPLIQUÉE AU TAM

La validité globale du TAM est discutée et évaluée à partir des propositions énoncées à la figure 2, laquelle présente les diverses facettes que peuvent prendre les validités spécifiques et les protocoles mis en place pour chacune des fonctions et assises concernées. La case A regroupe la validité de contenu, l'adéquation du

modèle de paramétrisation des items de la banque, la transformation des paramètres des items sur une même métrique selon la TRI, la vérification de l'efficience statistique du testing adaptatif selon une fonction d'information et la validité discriminante. À la case B appartiennent les règles de décision, l'établissement d'un critère de passage ou d'un score de césure, le choix des juges et l'application rigoureuse du protocole, l'évaluation globale du profil attendu de compétence scolaire. La case C comprend la validité corrélationnelle (reliée à des critères), le degré de concordance des décisions entre le testing adaptatif et le testing conventionnel, le respect des contraintes du milieu lors de l'expérimentation, le respect de la représentativité du programme d'études, le fonctionnement différencié entre les garçons et les filles, puis entre les initiés et les non-initiés à l'utilisation de micro-ordinateurs. La case D concerne la pertinence et la cohérence de la méthodologie utilisée en regard des objectifs de la recherche et de son contexte spécifique, ainsi que des décisions concernant les populations scolaires en regard des finalités du testing. Apprécier la validité globale du TAM, c'est donc prendre en compte les validités spécifiques et les protocoles méthodologiques en interaction entre eux et à l'intérieur d'une démarche évaluative.

ÉVALUATION DE LA VALIDITÉ GLOBALE D'UNE STRATÉGIE TAM APPLIQUÉE AU PROGRAMME D'ÉDUCATION ÉCONOMIQUE CINQUIÈME SECONDAIRE

Le programme d'éducation économique a été choisi parce qu'il partage les caractéristiques suivantes avec d'autres programmes en sciences humaines: être appliqué dans le milieu scolaire depuis un certain temps, offrir une certaine stabilité dans les résultats scolaires des élèves, offrir la possibilité d'utiliser des résultats à partir d'examens communs et standardisés, présenter la disponibilité de personnes-ressources du milieu scolaire oeuvrant dans le programme concerné. La figure 3 présente un résumé des contributions des facettes à la validité globale de la stratégie TAM telle qu'expérimentée et rapportée par Auger et Séguin (1992) selon les éléments A1 à D2 de la figure 2.

(A) Évidences empiriques: interprétation

La validité de contenu de la banque d'items a été établie par jugement d'experts sous la supervision des responsables de l'évaluation au ministère de l'Éducation du Québec (MÉQ). De plus, des analyses statistiques autant sous la théorie classique des scores que sous la TRI ont permis de ne conserver que les items valides du point de vue des évidences empiriques: les items non adéquats au modèle de Birnbaum (1968) ont été rejetés. Dans le but d'assurer une métrique commune à tous les items de la banque selon la TRI, une préexpérimentation a été conduite auprès d'élèves de cinquième secondaire en éducation économique: dix examens ont été administrés à une moyenne de 785 élèves par examen. Les examens avaient en moyenne 5 items communs sur une possibilité de 25 items

	Fonctions	
Assises	*Interprétation (validité interne)*	*Utilisation (validité externe)*
Évidences empiriques	A1: contenu validé par jugements d'experts. A2: analyses statistiques sous la TCS et la TRI et rejet des items non valides. A3: devis de 10 examens dont 20% d'items communs; technique à l'horizontale du ré-étalonnage des paramètres des items. A4: supériorité des quantités d'information au TAM par rapport au testing conventionnel. Réduction de 85% du nombre d'items administrés. A5: aucune donnée disponible.	C1: thêta au TAM et note d'école brute (r=0,61); thêta au TAM et note brute à l'examen du MÉQ (r=0,52). C2: degré d'accord entre TAM et MÉQ par rapport à la note d'école brute (50%) et à la note brute à l'examen (67%). C3: l'expérimentation s'est intégrée à l'horaire des cours de chacune des écoles participantes; la représentativité du contenu n'est pas assurée par le TAM. C4: aucune différence statistique au seuil de 0,05 au regard du fonctionnement différencié du TAM.
Conséquences des décisions	B1: identique à la pratique ministérielle pour le testing conventionnel. B2: application de la méthode d'Angoff au TAM et validité théorique de 0,95. B3: TAM et ses règles de décision vont dans le même sens que le profil attendu et défini par les experts en contenu.	D1: jugement en faveur de la pertinence et de la cohérence au regard des objectifs. D2: jugement d'adéquation du TAM en fonction d'un score total.

FIGURE 3

Évaluation de la validité globale au TAM appliquée en éducation économique cinquième secondaire

par examen. Ces items communs ont permis un ré-étalonnage uniforme des items selon une technique dite à l'horizontale et suivant une procédure de la moyenne et du sigma exploitant une relation linéaire bien connue (Hambleton et Swaminathan, 1985). L'efficience statistique du testing adaptatif (TA) a été établie par des rapports d'efficience relative, comparant les testing adaptatif et conventionnel en

termes de quantités d'information théorique I(Θ) à chaque niveau thêta. Les résultats concernant l'efficience du TA par rapport au testing conventionnel ont confirmé la supériorité de la stratégie TA. Cette supériorité a comme conséquence pratique de réduire de 85% le nombre d'items à être administrés tout en assurant le même degré de précision que le testing conventionnel basé sur une trentaine d'items. Comme nous ne disposions d'aucun critère externe fiable pour juger du classement des élèves, aucune analyse statistique de validité discriminante n'a toutefois été effectuée.

(B) Conséquences des décisions: interprétation

Les règles de décision tiennent compte de la pratique ministérielle en matière de certification scolaire. Le MÉQ déclare en situation de réussite un élève ayant un score égal ou supérieur à la note de passage de 60% plus ou moins 2% pour tenir compte de l'erreur de mesure soit, finalement, une note de passage de 58%. La même logique a guidé les règles de décision pour le testing adaptatif, avec la particularité d'être assujettie à la métrique habituelle de la TRI, soit un continuum entre –3 et +3. Le score de césure sous la TRI est le résultat de l'application rigoureuse de la méthode d'Angoff (1984). Cette méthode permet d'établir un score de césure en rapport avec une définition préalable de ce que doit démontrer un élève ayant une maîtrise suffisante des objets d'évaluation tout en se prononçant sur la probabilité que devrait avoir un tel élève de réussir chacun des items. Le score de césure est donc d'abord établi sur une base absolue pour devenir par la suite un critère-norme. La validité théorique du score de césure estimée en fonction du coefficient de fidélité inter-juges de la théorie de la généralisabilité (Cardinet et Tourneur, 1985) correspond à un coefficient maximum de 0,95. L'examen des profils des maîtrisants (M) et des non-maîtrisants (NM) à partir d'une épreuve unique contribue à apprécier la validité du score de césure pour l'ensemble des items de la banque ou du domaine mesuré.

Les résultats de l'étude démontrent une progression attendue de la moyenne de réussite par item selon un niveau d'habileté croissant (Auger et Séguin, 1992). Il a été démontré que les élèves classés M se distinguent de ceux classés NM quant à leur maîtrise ou non-maîtrise non seulement globale de l'ensemble du domaine mesuré, mais aussi de la majorité des sous-domaines considérés séparément.

(C) Évidences empiriques: utilisation

La validité reliée à des critères concerne le degré de corrélation entre les scores des mêmes élèves obtenus au TA et au testing conventionnel. Au moins deux critères retiennent l'attention soit la note d'école brute provenant des enseignants et la note brute obtenue à un examen unique du MÉQ tout à fait conforme à la représentativité du contenu. Les corrélations entre, d'une part, l'estimé d'habileté

au TAM, et, d'autre part, la note d'école brute, puis la note à l'examen unique, sont de l'ordre de 0,61 et de 0,52 respectivement. Les degrés d'accord entre le classement des élèves d'une part selon le TAM, qui est fonction du score de césure, et, d'autre part, le testing conventionnel qui est fonction de la note de passage, sont de l'ordre de 50% et 67% respectivement. En regard du fonctionnement différencié du TAM, aucune différence statistique au seuil de 0,05 n'a été trouvée entre les maîtrisants et non-maîtrisants, ni entre les initiés et les non-initiés aux micro-ordinateurs, sur quelque variable à l'étude.

Les contraintes du milieu ont été respectées par souci d'intégrer le TAM à partir de la planification des enseignants participant à l'expérimentation à l'intérieur de la grille-horaire hebdomadaire prévue par les établissements. La représentativité du contenu durant le TAM n'est pas confirmée, même si la banque d'items disponible au départ respecte la pondération proposée par le MÉQ. En réalité, ce résultat n'est pas surprenant, puisque le TAM cherche essentiellement à réduire l'erreur-type d'estimation autour de niveau individuel d'habileté. Pour les personnes soucieuses du respect de la représentativité du contenu, il y aura lieu de modifier la stratégie TAM tout en ne mettant pas en cause le bien fondé théorique du TAM.

(D) Conséquences des décisions: utilisation

La pertinence et la cohérence de la méthodologie utilisée en regard des objectifs et du contexte de la recherche sont soumises à une appréciation qualitative de l'ensemble de ces éléments pris comme un tout indissociable. Dans la perspective de vérifier l'efficience statistique de cette stratégie de testing adaptatif, et d'en examiner la praticabilité dans un contexte de sanction des études en vue d'une certification scolaire, il est maintenant possible d'affirmer que les choix reliés aux modèles théoriques d'estimation, ainsi que les protocoles méthodologiques mis en place, sont pertinents et cohérents pour l'atteinte des objectifs visés, particulièrement en ce qui regarde la validité interne. Pour l'aspect de la validité externe, quelques interrogations persistent quant aux degrés moyens de corrélation entre les variables critères et aux degrés de concordance des classements des élèves entre le TAM et le testing conventionnel. Toutefois, il ne faut pas oublier que le TAM est plus efficient que le testing conventionnel: jusqu'à quel point, alors, les valeurs moyennes de corrélation ou de concordance sont-elles attribuables aux limites du TAM ou à celles du testing conventionnel? Auger et Séguin (1992) ont indiqué en 16 points un devis efficace pour un testing adaptatif.

EN GUISE DE CONCLUSION

La réanalyse, sous le paradigme de la validité globale, des résultats d'une recherche déjà rapportée, a permis de circonscrire un nouveau cadre d'analyse tenant compte de la pratique usuelle d'une démarche évaluative. Cette volonté

d'intégrer les étapes d'un processus de démarche évaluative implique de considérer la validité globale comme un concept unificateur du processus d'évaluation. À travers la dynamique entre les validités spécifiques et les protocoles mis en place, il ressort que:
(a) le TAM possède une validité interne plus qu'intéressante autant du point de vue des preuves empiriques que des protocoles mis en place pour la détermination des règles de décision;
(b) le TAM fonctionne de manière uniforme autant pour les garçons que pour les filles ou que pour les initiés ou non à l'utilisation de l'ordinateur; et
(c) le TAM, tout en mesurant avec plus de précision les habiletés des élèves, ne classe pas ces derniers de la même manière qu'au testing conventionnel.

PROSPECTIVE D'ACTIVITÉS DE RECHERCHE ET DÉVELOPPEMENT

L'analyse de la validité globale de cette stratégie de testing adaptatif suggère quelques nouvelles recherches dans le domaine. Les résultats de ce TAM sont tributaires du modèle de Birnbaum. Ce modèle exige que près d'un millier d'individus soient testés pour fins de préexpérimentation des items. Un modèle moins exigeant en nombre de paramètres, comme le modèle de Rasch (1980), pourrait être aussi pertinent dans un contexte de certification scolaire. Une étude comparative entre les modèles de Birnbaum et de Rasch est présentement en cours. Il est anticipé que, pour une même précision de la mesure, il faudra augmenter le nombre d'items administrés en utilisant le modèle de Rasch et améliorer la représentativité du contenu.

Dans la préoccupation de réduction du nombre d'items à être administrés tout en assurant une efficience statistique égale pour tous les niveaux d'habileté et une représentation d'un profil attendu, la création d'une stratégie de testing adaptatif dans le domaine de la psychologie est aussi en cours (Tassé, 1994; Tassé, Maurice et Auger, 1993) soit une version informatisée et adaptative de l'Échelle Québécoise de Comportement Adaptatif (ÉQCA).

RÉFÉRENCES

American Psychological Association. (1985). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association/American Psychological Association/National Council of Measurement in Education.

Angoff, W. H. (1984). *Scales, norms and equivalent scores.* Princeton: N.J. Education Testing Service.

Auger, R. (1989). *Étude de praticabilité du testing adaptatif de maîtrise en regard des apprentissages scolaires au Québec: une expérimentation en éducation économique secondaire 5.* Thèse inédite de doctorat en éducation, Université du Québec à Montréal, Montréal, Québec.

Auger, R. et Dassa, C. (1992). Les pratiques de mesure et d'évaluation des apprentissages et la validité des tests dans un contexte de démarche évaluative. Dans D. Laveault et ADMÉE (Dirs.), *Les pratiques d'évaluation en éducation* (p. 151–166). Montréal: ADMÉE.

Auger, R. et Séguin, S. P. (1992). Le testing adaptatif avec interprétation critérielle, une expérience de praticabilité du TAM pour l'évaluation sommative des apprentissages au Québec. *Mesure et évaluation en éducation, 15*(1–2), 103–145.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. Dans F. M. Lord et M. Novick (Dirs.), *Statistical theories of mental test scores* (p. 397–479). Reading, MA: Addison Wesley.

Cardinet J. et Tourneur Y. (1985). *Assurer la mesure.* New York: Peter Lang.

Hambleton, R. K. et Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston, MA: Kluwer-Nighoff Publishing.

Hambleton, R. K., Swaminathan, H. et Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications.

Legendre, R. (1993). *Dictionnaire actuel de l'éducation* (2e éd.). Montréal: Guérin.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. Dans H. Wainer et H. I. Braun (Dirs.), *Test validity* (p. 33–45). Hillsdale, NJ: Lawrence Erlbaum Associates.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago Press.

Tassé, M. (1994). *Étude de la stabilité et de la concordance de l'Échelle québécoise de comportements adaptatifs (ÉQCA) et élaboration d'une version testage adaptatif informatisé de l'ÉQCA.* Thèse inédite de doctorat en psychologie, Université du Québec à Montréal, Montréal, Québec.

Tassé, M., Maurice, P. et Auger, R. (1993, mai). *Le testage adaptatif informatisé et l'évaluation des habiletés adaptatives.* Communication présentée au colloque de la Technologie informatique au profit des personnes en difficulté, CÉGEP du Vieux-Montréal, Montréal, Québec.

Wainer, H. et Braun, H. I. (Dirs.). (1988). *Test validity.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J. et Steinberg, L. (Dirs.). (1990). *Computerized adaptive testing: a primer.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Weiss, D. J. et Vale, C. D. (1987). Adaptive testing. *Applied Psychology: An International Review, 36*(3–4), 249–262.

Réjean Auger et Serge P. Séguin sont professeurs au Département des sciences de l'éducation, l'Université du Québec à Montréal, Case postale 8888, succursale centre-ville, Montréal (Québec) H3C 3P8.