Titre abrégé: MODÈLES DE MESURE EN ÉDUCATION

Les modèles de mesure en éducation : enjeux, développements et orientations

Pierre Valois[1]
Université Laval

Romain Martin[2]
Université du Luxembourg

Address : [1] Pierre Valois[1]
Université Laval
Département des fondements et pratiques en éducation
Local 462 TSE, Université Laval
2320, rue des bibliothèques
Québec (Québec) G1V 0A6 Canada
E-mail : Pierre.Valois@fse.ulaval.ca

[2] Romain Martin
Université du Luxembourg
Faculté LSHASE - Campus Walferdange
Château de Walferdange
Bâtiment XII, rdch, bureau 0.2
Route de Diekirch
L-7220 Walferdange
Luxembourg / Europe
E-mail : romain.martin@uni.lu

Résumé

Cet article vise à dresser un bilan des études ayant porté sur les modèles de mesure ou sur un aspect particulier de ces derniers au cours des 30 dernières années. Pour ce faire, nous avons réalisé une recension des écrits à partir de 14 revues internationales traitant, entre autres, des modèles de mesure en éducation. Après avoir effectué une synthèse des résultats qui ressortent de notre analyse des articles sélectionnés, nous présentons des développements et des défis majeurs qui s'imposent encore aujourd'hui à ceux et celles qui travaillent à l'élaboration, au raffinement et à l'utilisation des modèles de réponse aux items. Enfin, dans un dernier temps, nous proposons un certain nombre d'avenues de recherche pour les années à venir.

Mots-clés : modèles de réponses aux items, modèles de mesure, testing adaptatif par ordinateur, mesure de tâches complexes, enquêtes à large échelle

Abstract

This article presents a systematic review of literature on measurement models, and in particular on Item Response Models, during the last 30 years. First, we conduct a review of articles selected from fourteen international journals on measurement in education. Second, we present the results of this review of literature and the developments and the major challenges that remain to be tackled by researchers working on elaborating, refining and using Item Response Models. Finally, we suggest new avenues to be explored in the coming years.

Key Words: Item response models, measurement models, computerized adaptive testing, scoring of complex tasks, large-scale surveys

La question de la mesure constitue encore à ce jour un sujet de discorde entre les chercheurs. Les discussions et les prises de position entourant l'utilité réelle de la mesure (par exemple, la mesure des apprentissages ou des acquis d'expérience) baignent souvent dans un climat où les principes établis et les données empiriques s'emmêlent à des idées préconçues sur le sujet. Dans le domaine de l'éducation, puisque c'est celui qui nous intéresse d'une façon plus particulière, un certain nombre d'auteurs et de professionnels affirment sans détour ne pas croire, ou encore bien peu, à l'utilité des théories de la mesure en usage pour appuyer la construction de dispositifs d'évaluation. Ainsi, par exemple et pour ne citer que ceux-là, Bercier-Larivière et Forgette-Giroux (1999), soutiennent que « le modèle psychométrique, avec ses critères de validité et de fidélité, ne convient pas au domaine scolaire » (p. 169).

En clair, certains auteurs soutiennent que c'est de faire fausse route que de prendre appui sur les théories psychométriques pour construire et valider les outils d'évaluation des apprentissages. Pour d'autres, ces prises de position étonnent puisque, de leur point de vue, les principes et techniques mises de l'avant par la psychométrie et les théoriciens de la mesure ne visent qu'à favoriser le respect de critères inhérents à la nature même de la mesure. Si le recours à des modèles inspirés de la psychométrie est aussi fréquent dans le monde de l'éducation, cela tiendrait d'abord au fait que, de par leur nature, les propriétés mesurées s'apparentent davantage à celles rencontrées dans le domaine de la psychologie que dans celui des sciences exactes où les caractéristiques mesurées prennent souvent appui une unité standard de référence, une unité formelle de mesure. Si les théories de la mesure sont ce qu'elles sont dans nos domaines de recherche et d'activité, cela s'expliquerait principalement par la nature même des construits en jeu ainsi qu'à la complexité de leur évaluation.

Le rejet des apports des théories de la mesure ou la négation simple de leur utilité étonnent néanmoins du fait que cette attitude apparaît absolument dissonante par rapport à celle des associations professionnelles les plus importantes dans le domaines de l'éducation (*National Council on Measurement in Education, American Evaluation Association, American Federation of Teachers, Canadian Society for the Study of Education* et plusieurs autres) qui s'efforcent précisément de définir des standards d'évaluation des étudiants prenant appui sur les avancés de ces théories, que ces dernières soient issues de la psychologie, de la sociologie ou autres (The Joint Committee on Standards for Educational Evaluation, 2003).

Sans doute est-on trop souvent avare de nuance aussi bien lorsqu'il s'agit d'appliquer à une réalité quelque peu incertaine quelque principe ou règle des théories de la mesure que dans la dénégation de la valeur de ces dernières sur la seule base de limites imposées par la nature de la réalité considérée. Pour justifier leur rejet les auteurs traitent souvent de manière très superficielle des sujets pourtant complexes et enfoncent sans arrêt le clou des limites potentielles et réelles de ces outils. En revanche, des tenants de la position contraire présentent parfois comme des absolus des « arguments » portant discutables, mettent la pédale douce sur les limites des modèles qu'ils exploitent ou encore oublient de garder à l'esprit l'objectif simple et parfois rudimentaire de certaines mesures avant de proposer ou de s'élancer dans des approches inutilement complexes et exagérément onéreuses compte tenu du faible niveau de raffinement des informations recherchées. À un niveau d'observation donné, tout avancé théorique se heurte à quelques limites. De telle sorte que la posture à la fois la plus judicieuse et la plus confortable se situe probablement quelque part entre celles des tenants des positions extrêmes sur le sujet.

Le présent texte vise d'abord à dresser un bilan des études ayant porté sur l'un ou l'autre des modèles de mesure ou sur un aspect particulier de ces derniers au cours des 30 dernières

années. Pourquoi 30 ans ? Parce que cela correspond à l'année de parution du premier numéro de la revue *Mesure et évaluation en éducation*, soit en 1978. Nous ne prétendons pas que cette recension soit exhaustive. Elle devrait néanmoins permettre aux personnes intéressées ou concernées par la mesure et l'évaluation d'avoir une assez juste idée des avancés dans le domaine, des enjeux qui s'y jouent ainsi que des développements majeurs qui caractérisent ce secteur de recherche aujourd'hui. Comme il se doit, il sera, à travers cette recension des écrits, aussi question de la place réservée par la revue Mesure et évaluation en éducation aux différents modèles de mesure depuis sa première parution.

Cet article comporte quatre sections principales. Dans un premier temps, nous rendrons compte dans le détail de la méthodologie utilisée pour répertorier les articles d'intérêt et parvenir à identifier les enjeux et développements majeurs actuels au sujet des modèles de mesure. Une fois décrite cette démarche ou ce mode d'opération, nous tracerons, dans un deuxième temps, les grandes lignes des résultats qui se dégagent de notre analyse des articles sélectionnées. Une troisième section sera consacrée à la présentation des développements principaux ainsi que des défis majeurs qui s'imposent encore aujourd'hui à ceux et celles qui travaillent à l'élaboration et au raffinement des modèles de mesure. Enfin, dans un dernier temps, nous tenterons d'identifier un certain nombre des avenues de recherche les plus sollicitées pour les années à venir et discuterons brièvement des orientations que pourrait ou devrait, selon les points de vue, prendre la revue *Mesure et évaluation en éducation* au regard du sujet qui nous intéresse.

Nous sommes absolument conscients que tous ne partageront pas notre lecture de la réalité et nos perceptions. Ce que nous considérons un enjeu, un développement majeur ou une avenue à développer pourra, aux yeux de certains, s'apparenter davantage à un ensemble de thèmes sans véritable importance. Par ailleurs, nous reconnaissons d'emblée qu'on puisse

différer d'opinion sur la pertinence des modèles de mesure, leur importance ou leur utilité relative. Ce qui importe avant tout, c'est de s'ouvrir à l'éventail des points de vue, d'accepter de confronter objectivement et dans le détail nos divergences et d'y voir une réelle opportunité de progresser, de bonifier l'ensemble des connaissances sur le sujet.

Méthode

Constitution de l'échantillon d'articles

Cette démarche comprend le choix des revues scientifiques. Attardons-nous d'abord au choix des revues. Parmi toutes les sources possibles d'articles relatifs aux modèles de mesure, nous avons retenu, sur la base leur notoriété à titre de sources de publications dans le domaine, 14 revues scientifiques. Plus précisément, 13 des 14 revues retenues sont anglophones, soit: (1) *Applied Measurement in Education,* (2) *Applied Psychological Measurement,* (3) *Educational and Psychological Measurement,* (4) *Journal of Educational Measurement,* (5) *Measurement and Evaluation in Counseling and Development,* (6) *Psychometrika,* (7) *Quality & Quantity,* (8) *Educational Measurement: Issues and Practice,* (9) *International Journal of Testing,* (10) *Measurement: Interdisciplinary Research and Perspective,* (11) *Measurement in Physical Education and Exercise Science,* (12) *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences,* (13) *Studies in Educational Evaluation.* Du côté francophone, nous avons retenu la revue *Mesure et évaluation en éducation*. Selon le Journal of Citation Reports, 7 des 13 revues anglophones ont une cote d'impact variant entre 0,225 et 0,792 (i.e. les 7 premières revues de la liste précédente). Quant aux 6 autres revues, ils constituent un choix personnel basé sur notre connaissance des écrits scientifiques sur les modèles de mesure.

En ce qui concerne maintenant le choix des articles, nous avons restreint notre sélection à ceux dont l'objet principal concernait directement le développement d'un ou de modèles de mesure en éducation, et ce, pour deux raisons. Premièrement, parce que l'objectif du présent article consiste principalement à dresser le portrait du développement des modèles de mesure et non pas à en faire une description. Deuxièmement, si nous avions recensé tous les articles qui ont rapporté avoir eu recours à l'un ou l'autre de ces modèles sans avoir comme objet d'études leur développement théorique, nous nous serions alors retrouvés avec un nombre considérable d'articles desquels il aurait été difficile de dégager des conclusions. À titre d'exemple, il aurait été tout à fait insensé, de vouloir recenser tous les articles ayant eu recours au coefficient alpha de Cronbach pour éprouver la fidélité de leur instrument de mesure.

Après avoir procédé au choix des sources de publication, nous avons, de manière systématique, recensé les articles ayant traité des modèles de mesure ou encore ayant fait appel à eux. Dans le cas des revues anglophones, nous avons effectué la recension à l'aide des principales bases de données Internet, dont S*pringerlink, CSA* et *EBSCO,* pour ne nommer que celles-ci. Une première recherche fut d'abord effectuée sur les résumés des articles à l'aide des mots clés suivants : « item response theory », « nonparametric item response theory », « generalisability theory » et « classical test theory »[1].

Classement des articles

L'échantillon d'articles a été ensuite réparti sur la base de trois critères, le premier étant la décennie de publication : (1) 1978 à 1987, (2) 1988 à 1997 et (3) 1998 à 2008. Le second critère est le modèle de mesure analysé ou utilisé dans l'article et il se décline en quatre catégories : (1) la théorie classique des tests ou TCT, (2) la théorie de la généralisabilité ou TG

[1] Nous sommes conscients que nous aurions pu utiliser d'autres mots clés tel que « structural equation modelling », « neural model » ou « multicomponent latent trait model », mais nous avons jugé nécessaire de se limiter aux modèles découlant de ces quatre théories pour mieux circonscrire notre recension des écrits.

(3) les modèles paramétriques de la théorie des réponses aux items ou TRI et (4) les modèles non paramétriques (NP) de la théorie des réponses aux items. Enfin, le thème central de l'article constitue le troisième critère de sélection. Six thèmes nous sont apparus particulièrement intéressants compte tenu des défis actuels dans le domaine de la mesure et l'évaluation en éducation, soit : (1) les articles d'ordre théorique et mathématique; (2) le calibrage des items; (3) le testing adaptatif; (4) la dimensionalité des tests; (5) le fonctionnement différentiel d'items (FDI); (6) l'évaluation automatique des tâches complexes.

En résumé, les articles des 14 revues analysées ont été répartis dans une matrice 3 × 4 × 6 reflétant le croisement des trois variables de classification, c'est-à-dire la décennie de publication, le modèle de mesure analysé ou utilisé et le thème central de l'article. La procédure de classement des articles a toutefois été quelque peu adaptée pour la revue Mesure et évaluation en éducation. Nous reviendrons plus loin dans l'article sur ce point.

Résultats de la recension des articles

Cette section se divise en deux parties. Nous allons d'abord présenter les résultats quantitatifs de la recension des articles provenant des revues internationales anglophones. Par la suite, nous dégagerons les contributions spécifiques à la revue *Mesure et évaluation en éducation*.

Articles provenant des revues internationales anglophones

Notre recherche informatisée des articles publiés dans des revues internationales anglophones a débouché sur l'identification de 621 articles traitant directement des modèles de mesure en éducation. L'examen de la dernière colonne du tableau 1 indique que ce sont les revues *Applied Psychological Measurement* (36,7 %), *Journal of Educational Measurement*

(16,7 %) et *Educational and Psychological Measurement* (15,6 %) qui publient le plus d'articles sur les modèles de mesure.

Le tableau 2 a été élaboré afin de mieux faire ressortir l'importance relative de chacun des thèmes de recherche et des modèles de mesure. L'examen des cellules qui correspondent aux croisements de la dernière ligne avec les colonnes identifiant les décades (voir le tableau 2) indique que la parution d'articles portant sur les modèles de mesure a augmenté continuellement au cours des 3 dernières décades. En effet, on peut remarquer que le nombre d'articles publiés est passé de 71 (11,4 %) à la première décade (1978-1987), à 216 (34,8 %) à la seconde (1988-1997) et enfin à 334 (53,8 %) entre 1998 et 2008. Des 621 articles, la majorité concerne les modèles de la TRI : n=444 (71,5 %;). Pour les autres catégories, la répartition est la suivante : TCT : n=55 (8,9 %); TG : n=82 (13,2%); non paramétrique : n=40 (6,4 %).

Une analyse plus détaillée en fonction des variables de répartition indique que sur les 55 articles ayant traité de la théorie classique des tests (TCT), 8 (14,5 %) l'ont été lors de la première décade, 25 (45,5 %) lors de la 2^{e} décade et 22 (40,0 %) après 1998. En ce qui a trait aux 82 articles de la théorie de la généralisabilité (TG), ils se déclinent comme suit : 20 (24,4 %), 16 (19,5%) et 46 (56,1%). Pour les modèles de la théorie de la réponse aux items (TRI), le nombre de publications augmentent considérablement d'une décade à une autre : 40 (9 %), 160 (36 %) et 244 (55 %). Enfin, bien que le nombre de publications pour les modèles non paramétriques (NP) soit plutôt faible, il augmente quand même d'une décade à une autre : 3 (7,5 %), 15 (37,5 %) et 22 (55,0 %).

On constate donc que le nombre d'articles portant sur les modèles de la TCT a augmenté entre 1978-1987 et 1988-1997 pour se stabiliser par la suite. La courbe d'augmentation est différente pour les articles portant sur la généralisabilité (TG). Bien que la théorie ait pris racine

il y a déjà plusieurs années, le nombre d'articles publiés a augmenté lors de la dernière décade, cette progression étant sans doute due à l'utilité pour évaluer la fidélité des nouveaux dispositifs d'évaluation dans le cadre des programmes de formation élaborés à partir de la notion de compétence. Ces modèles s'avèrent particulièrement utiles pour déterminer le nombre d'évaluateurs ou de tâches requises pour assurer la fidélité des dispositifs d'évaluation déployés. Par ailleurs, la TG possède d'autres domaines d'application plus récents. Par exemple, dans un récent numéro thématique de la revue Mesure et évaluation en éducation consacré au thème de la généralisabilité, Cardinet (2003) présente une démarche qui s'appuie qui sur la TG pour certifier les progrès des élèves de façon individualisée.

Quant aux modèles de la TRI, les données indiquent qu'ils se sont considérablement développés depuis dix ans. Ces divers modèles permettent désormais aux chercheurs et professionnels de la recherche de recourir à des modèles de réponse aux items plus performants pour mieux développer et valider leur outils d'évaluation, qu'il s'agisse de questionnaires d'attitude, de motivation, d'examens ou des questionnaires développés à des fins de sélection de personnel, la performance au travail, l'évaluation des troubles de comportements des élèves, etc.

Par ailleurs, à la lumière des fréquences présentées dans les trois colonnes correspondant à la TRI (tableau 3), il appert que la majorité des articles publiés sont d'ordre théorique et mathématique. Ceci n'est pas surprenant compte tenu que les modèles de mesure basés sur la TRI sont encore en plein expansion et que les chercheurs s'y intéressant sont de plus en plus nombreux et proviennent, par surcroît, de plusieurs sphères de recherche.

Articles provenant de la revue *Mesure et évaluation en éducation*

En ce qui a trait à la revue *Mesure et évaluation en éducation*, nous avons fait preuve d'une certaine souplesse quant aux choix de nos critères de sélection des articles. En effet,

compte tenu de la mission élargie de la revue au regard du type de recherche qu'elle publie et du vivier relativement restreint d'articles sur les modèles de mesure, nous avons jugé opportun de sélectionner non seulement les articles théoriques et empiriques, mais aussi ceux à caractères plus pédagogiques (i.e. descriptif ou réflexif) de même que ceux dont le recours aux modèles de mesure ne visait qu'à estimer la validité de leur instrument de mesure (i.e. coefficients alpha de Cronbach ou de généralisabilité). Compte tenu de cette relative souplesse dont nous avons fait preuve dans le choix des articles, il est quand même surprenant de constater que sur les 582 articles publiés dans la revue depuis 30 ans, seulement 63 (10,8 %) ont porté sur les modèles de mesure.

L'examen du tableau 3 permet de constater que de ces 63 articles, 29 (46 %) réfèrent directement à la théorie classique des tests. Toutefois, pour 24 de ces 29 articles c'est le recours au coefficient alpha de Cronbach qui a justifié notre décision de les considérer dans la recension. Le tableau 3 indique aussi que 26 articles (41,3 %) traitent de la TRI, 8 (12,7 %) de la TG et aucun des modèles non paramétriques. Un examen plus détaillé des résultats démontrent que depuis 1988 le nombre d'articles s'étant intéressés aux modèles des réponses aux items a augmenté considérablement : 0 entre 1978-1987, 16 entre 1988-1997 et 9 entre 1998 et 2008, ce qui n'est pas le cas pour les trois autres théories. De ces 27 articles, 9 (33 %) sont d'ordre descriptif ou réflexif (p. ex. Blais & Ajar, 1991; Loye, 2005); 8 (29,6 %) portent sur le testing adaptatif (p. ex. Auger, 1992; Laurier, 1996), 4 (14,3 %) sur le calibrage d'items et l'estimation des habiletés et des paramètres d'items (p. ex. Burton,), 2 sur la détermination du nombre de dimension dans un test (Blais & Laurier, 1997; Raîche, Langevin, Riopel, & Mauffette, 2006), 2 sur le fonctionnement différentiel d'items (Bertrand, 2001**;** Dechef & Laveault, 1993). Enfin aucun article dans notre recension n'a porté directement sur les enquêtes internationales, ni sur

l'évaluation automatique des tâches complexes. L'article de Lafontaine et Simon (2008) dans ce numéro aborde cependant les utilisations des modèles de la TRI dans les enquêtes internationales visant l'évaluation des systèmes éducatifs.

Il ressort de l'examen de cette recension des écrits que peu d'articles traitant des modèles de mesure ont été publiés de la revue Mesure et évaluation en éducation depuis sa parution en 1978. De plus, la plupart de ces articles sont des descriptions des modèles de mesure ou des réflexions sur ceux-ci. Enfin, dans l'ensemble les deux recensions des écrits font ressortir l'intérêt croissant des chercheurs pour les modèles de mesure de la TRI et ce particulièrement depuis le milieu des années 80. La prochaine section sera consacrée à la présentation des enjeux et développements majeurs relatifs aux modèles de réponses aux items (MRI).

Modèles de réponse aux items : enjeux et développements majeurs

Parallèlement à l'accroissement des activités de recherche en évaluation des acquis scolaires, des établissements ou encore à des fins de reconnaissance des acquis d'expérience, pour ne nommer que ces secteurs d'intervention, se manifeste une préoccupation constante au sujet de la qualité des outils d'évaluation utilisés. Disposer d'instruments de mesure de bonne qualité, renvoyant des informations à la fois pertinentes, précises et à l'abri de biais susceptibles de léser un ou des sous-groupes quelconques de ceux à qui il est administré, voilà le vœu de tous mais aussi un des défis auxquels sont confrontés tant les chercheurs que les intervenants œuvrant dans les secteurs de l'évaluation des apprentissages, des établissements scolaires, des systèmes et des politiques de l'éducation, etc. À plusieurs d'entre eux, se pose également le problème du calibrage des items tirés de différents questionnaires présumés mesurer le même construit, des patrons aberrants de réponse (St-Onge, Valois, Abdous, & Germain, in press) ou encore de choisir les meilleurs items pour identifier avec précision les répondants situés au voisinage d'un seuil ou d'un point de

césure donné (*cutoff score*). Il s'agit là de quelques exemples seulement de l'immense bassin des problèmes complexes auxquels on peut être confronté. Cet état de fait s'explique d'une part par la complexité des construits et des concepts étudiés en éducation et d'autre part par le recours à des modèles de mesure de plus en plus complexes.

Tableau 1. Articles sur les modèles de mesure publiés dans des revues internationales en fonction des années, des théories et des thèmes de recherche

Périodiques	Thème	1978-1987					1988-1997					1998-2008					Grand total	
		TCT	TG	TRI	NP	Total 1e décade	TCT	TG	TRI	NP	Total 2e décade	TCT	TG	TRI	NP	Total 3e décade		
Applied Measurementin Education (1988-2008)	Théorique / modèles mathématiques					0	1	2	15	2	20	1	6	5		12	32	
	Calibrage					0			5		5			5		5	10	
	Testing adaptatif					0					0			2		2	2	
	Dimensionalité d'un test					0					0					0	0	
	Fonctionnement différentiel d'items (FDI)					0			3		3			1		1	4	
	Évaluation automatique des tâches complexes					0					0					0	0	
	Total	0	0	0	0	0	1	2	23	2	28	1	6	13	0	20	48	7,7%
Psychometrika (1978-2008)	Théorique / modèles mathématique	1		2	1	4	2	1	9	4	16	4		19	2	25	45	
	Calibrage					0					0					0	0	
	Testing adaptatif					0					0			2		2	2	
	Dimensionalité d'un test			2	1	3			1		1			5		5	9	
	Fonctionnement différentiel d'items (FDI)					0			1		1					0	1	
	Évaluation automatique des tâches complexes					0					0					0	0	
	Total	1	0	4	2	7	2	1	11	4	18	4	0	26	2	32	57	9,2%
Applied Psychological Measurement (1978-2008)	Théorique / modèles mathématiques	6	5	11	1	23	8		33	6	47	6	2	40	9	57	127	
	Calibrage			7		7			8		8			19		19	34	
	Testing adaptatif			1		1	1		7		8			14		14	23	
	Dimensionalité d'un test	1				1	2		3	1	6			8	6	14	21	
	Fonctionnement différentiel d'items (FDI)			2		2			16		16			5		5	23	
	Évaluation automatique des tâches complexes					0					0					0	0	
	Total	7	5	21	1	34	11	0	67	7	85	6	2	86	15	109	228	36,7%
Journal of Educational Measurement (1978-2008)	Théorique / modèles mathématiques		5	7		12	4	4	14		22	1	9	22	2	34	68	
	Calibrage			2		2			7		7	1		2		3	12	
	Testing adaptatif			1		1			3		3			1	1	2	6	
	Dimensionalité d'un test		1			1			3		3			4		4	8	
	Fonctionnement différentiel d'items (FDI)					0			5		5			5		5	10	
	Évaluation automatique des tâches complexes					0					0					0	0	
	Total	0	6	10	0	16	4	4	32	0	40	2	9	34	3	48	104	16,7%

Périodiques	Thème	1978-1987 TCT	TG	TRI	NP	Total 1e décade	1988-1997 TCT	TG	TRI	NP	Total 2e décade	1998-2008 TCT	TG	TRI	NP	Total 3e décade	Grand total	
Educational and Psychological Measurement (1978-2008)	Théorique / modèles mathématiques		7			7	5	4	3	2	14	3	11	28		42	63	
	Calibrage					0					0			3		3	3	
	Testing adaptatif					0			1		1			1		1	2	
	Dimensionalité d'un test					0		1	7		8	1	2	9		12	20	
	Fonctionnement différentiel d'items (FDI)					0			3		3	1		5		6	9	
	Évaluation automatique des tâches complexes					0					0					0	0	
	Total	0	7	0	0	7	5	5	14	2	26	5	13	46	0	64	97	15,6%
International Journal of Testing (2001-2007)	Théorique / modèles mathématiques					0					0		1	9	2	12	12	
	Calibrage					0					0			2		2	2	
	Testing adaptatif					0					0			1		1	1	
	Dimensionalité d'un test					0					0		1	1		2	2	
	Fonctionnement différentiel d'items (FDI)					0					0			1		1	1	
	Évaluation automatique des tâches complexes					0					0					0	0	
	Total	0	0	0	0	0	0	0	0	0	0	0	2	14	2	18	18	2,9%
Educational Measurement: Issues and Practice (1982-2008)	Théorique / modèles mathématiques		1	4		5	1	2	6		9	3	4	6		13	27	
	Calibrage					0			1		1					0	1	
	Testing adaptatif					0			2		2			1		1	3	
	Dimensionalité d'un test					0					0		1	1		2	2	
	Fonctionnement différentiel d'items (FDI)					0			1		1			1		1	2	
	Évaluation automatique des tâches complexes					0					0					0	0	
	Total	0	1	4	0	5	1	2	10	0	13	3	5	9	0	17	35	5,6%
Measurement: Interdisciplinary Research & Perspective (2003-2008)	Théorique / modèles mathématiques					0					0		1	2		3	3	
	Calibrage					0					0					0	0	
	Testing adaptatif					0					0					0	0	
	Dimensionalité d'un test					0					0			5		5	5	
	Fonctionnement différentiel d'items (FDI)					0					0					0	0	
	Évaluation automatique des tâches complexes					0					0					0	0	
	Total	0	0	0	0	0	0	0	0	0	0	0	1	7	0	8	8	1,3%
Measurement in physical education and Exercise Science (1997-2008)	Théorique / modèles mathématiques					0	1				1		2			2	3	
	Calibrage					0					0					0	0	
	Testing adaptatif					0					0					0	0	
	Dimensionalité d'un test					0					0		3	1		4	4	
	Fonctionnement différentiel d'items (FDI)					0					0					0	0	
	Évaluation automatique des tâches complexes					0					0					0	0	
	Total	0	0	0	0	0	1	0	0	0	1	0	5	1	0	6	7	1,1%

Périodiques	Thème	1978-1987 TCT	TG	TRI	NP	Total 1e décade	1988-1997 TCT	TG	TRI	NP	Total 2e décade	1998-2008 TCT	TG	TRI	NP	Total 3e décade	Grand total	
Methodology: European Journal of Research Methods for the Behavioral and Social Sciences (2005-2008)	Théorique / modèles mathématiques					0					0			3		3	3	
	Calibrage					0					0			1		1	1	
	Testing adaptatif					0					0					0	0	
	Dimensionalité d'un test					0					0					0	0	
	Fonctionnement différentiel d'items (FDI)					0					0					0	0	
	Évaluation automatique des tâches complexes					0					0					0	0	
	Total	0	0	0	0	0	0	0	0	0	0	0	0	4	0	4	4	0,6%
Studies in Educational Evaluation (1978-2008)	Théorique / modèles mathématiques			1		1			1		1		1	1		2	4	
	Calibrage					0					0					0	0	
	Testing adaptatif					0					0					0	0	
	Dimensionalité d'un test					0					0					0	0	
	Fonctionnement différentiel d'items (FDI)					0					0					0	0	
	Évaluation automatique des tâches complexes					0					0					0	0	
	Total	0	0	1	0	1	0	0	1	0	1	0	1	1	0	2	4	0,6%
Quality & Quantity (1978-2008)	Théorique / modèles mathématiques		1			1		2			2	1	1	2		4	7	
	Calibrage					0					0					0	0	
	Testing adaptatif					0					0					0	0	
	Dimensionalité d'un test					0					0					0	0	
	Fonctionnement différentiel d'items (FDI)					0					0					0	0	
	Évaluation automatique des tâches complexes					0					0					0	0	
	Total	0	1	0	0	1	0	2	0	0	2	1	1	2	0	4	7	1,1%
Measurement and Evaluation in Counseling and Development (1990-2008)	Théorique / modèles mathématiques					0			1		1		1			1	2	
	Calibrage					0					0					0	0	
	Testing adaptatif					0			1		1			1		1	2	
	Dimensionalité d'un test					0					0					0	0	
	Fonctionnement différentiel d'items (FDI)					0					0					0	0	
	Évaluation automatique des tâches complexes					0					0					0	0	
	Total	0	0	0	0	0	0	0	2	0	2	0	1	1	0	2	4	0,6%
	Grand total	8	20	40	3	71	25	16	160	15	216	22	46	244	22	334	621	
	%	11,3%	28,2%	56,3%	4,2%		11,6%	7,4%	74,1%	6,9%		6,6%	13,8%	73,1%	6,6%			

Tableau 2. Fréquences des thèmes de recherche abordés dans les articles portant sur les modèles de mesure et publiés dans des revues internationales en fonction des années et des théories de la mesure

Thème	1978-1987					1988-1997					1998-2008					Grand total	%
	TCT	**TG**	**TRI**	**NP**	Total 1re décade	**TCT**	**TG**	**TRI**	**NP**	Total 2e décade	**TCT**	**TG**	**TRI**	**NP**	Total 3e décade		
Théorique / modèles mathématiques	7	19	25	2	53	22	15	82	14	133	19	39	137	15	210	396	63,8%
Calibrage	1	1	2	1	5	2	1	14	1	18	1	7	34	6	48	71	11,4%
Testing adaptatif	0	0	9	0	9	0	0	21	0	21	1	0	32	0	33	63	10,1%
Dimensionalité d'un test	0	0	2	0	2	1	0	14	0	15	0	0	23	1	24	41	6,6%
Fonctionnement différentiel d'items (FDI)	0	0	2	0	2	0	0	29	0	29	1	0	18	0	19	50	8,1%
Évaluation automatique des tâches complexes	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,0%
Total	8	20	40	3	71	25	16	160	15	216	22	46	244	22	334	621	
%	11,3%	28,2%	56,3%	4,2%		11,6%	7,4%	74,1%	6,9%		6,6%	13,8%	73,1%	6,6%			

Tableau 3. Articles portant sur les modèles de mesure et publiés dans la revue Mesure et évaluation en éducation en fonction des années, des théories de la mesure et des thèmes de recherche

Thèmes	1997-1987					1988-1997					1998-2008					Grand total	%
	TCT	**TG**	**TRI**	**NP**	Total 1re décade	**TCT**	**TG**	**TRI**	**NP**	Total 2e décade	**TCT**	**TG**	**TRI**	**NP**	Total 3e décade		
Théorique					0	3				3	1				1	4	6,3%
Descriptif / Réflexif					0	1	2	8		11		4	1		5	16	25,4%
Fidélité (coeff. alpha, coeff. de généralisabilité)					0	12				12	12	2			14	26	41,3%
Calibrage					0			1		1			3		3	4	6,3%
Testing adaptatif			1		1			5		5			3		3	9	14,3%
Dimensionalité d'un test					0			1		1			1		1	2	3,2%
Fonctionnement différentiel d'items (FDI)					0			1		1			1		1	2	3,2%
Enquêtes internationales					0					0					0	0	0%
Évaluation automatique des tâches complexes					0					0					0	0	0%
Total	0	0	1	0	1	16	2	16	0	34	13	6	9	0	28	63	
%	0%	0%	100%	0%		47,1%	5,8%	47,1%	0%		46,4%	21,4%	32,2%	0%			

Les données des tableaux précédents illustrent le fait, qu'au cours des dernières années, ces réflexions ont suscité beaucoup d'intérêt chez les chercheurs, intérêt qui s'est traduit par un nombre impressionnant de publications scientifiques ainsi que l'organisation de plus d'une dizaine de congrès et conférences internationales spécifiquement consacrés à la question des modèles de mesure (p. ex., l'*International Conference in Rasch Measurement*). Les problématiques abordées sont très variées et reflètent l'omniprésence de la synergie entre les domaines de la psychométrie, la statistique et l'éducation qui a favorisée le développement de nouvelles stratégies d'élaboration et de validation d'outils de mesure basés sur les théories modernes de mesure, notamment la TRI.

En bref, dans le but d'évaluer des résultats tels que les scores des élèves, la performance des établissements scolaires ou le rang des pays aux enquêtes internationales, à l'aide d'instruments de mesure fiables, valides et interprétables, un effort soutenu est mis sur le développement de nouvelles approches ou la révision et l'amélioration d'outils déjà existants. Ainsi, l'adoption de la théorie de la réponse aux items (TRI) au détriment de la théorie classique des tests (TCT) est de plus en plus courante. Ce succès de la TRI est dû, entre autres, à sa capacité de réaliser l'analyse d'échelles et d'items, de calibrer des instruments différents et de réaliser des tests adaptatifs.

Toutefois, les MRI dépendent – pour devenir pleinement opérationnels – de deux conditions majeurs : disposer d'un nombre de sujets suffisant afin de pouvoir appliquer les procédures d'étalonnage et disposer d'une puissance de calcul suffisante afin de mettre en œuvre les algorithmes complexes qui font partie intégrante des MRI d'une manière efficace et rapide. Ces deux conditions ont été réalisés dans les dernières années à travers les enquêtes nationales et notamment internationales à large échelle et à travers le développement de l'infrastructure

informatique qui favorise la réalisation du testing adaptatif par ordinateur. Afin de comprendre dans quelle mesure les MRI ont constitué un modèle de mesure particulièrement adapté à la fois aux enquêtes à large échelle et au testing adaptatif, il faut se rappeler les caractéristiques principales des MRI, ainsi que les finalités qui sont rattachées d'un côté aux enquêtes à large échelle et de l'autre côté au testing adaptatif. La caractéristique principale qui rend les MRI particulièrement intéressants pour les deux cas d'application mentionnés est le fait que ce modèle de mesure permet de situer les sujets évalués sur une même dimension de compétence à partir de résultats obtenus sur des échantillons d'items différents. En d'autres mots, si on dispose d'un ensemble d'items étalonnés à l'aide du MRI, même si les sujets passent des tests qui ne sont pas identiques, il est néanmoins possible de calculer pour ces sujets des paramètres de compétence qui les placent tous sur la même dimension de compétence qui est d'ailleurs identique à la dimension de difficulté sur laquelle les items ont été positionnés suite à leur étalonnage.

Enquêtes à large échelle

Pour les enquêtes comparatives internationales à large échelle, telle que le PISA ou TIMSS, le but principal est d'arriver à une estimation précise des paramètres de compétence (moyennes, variance, répartition à travers des niveaux de compétences différents, percentiles, erreur-standard) pour des populations et des sous-populations. Le grand souci de ce genre d'enquête est de pouvoir comparer des (sous-)populations différentes sur les mêmes dimensions de compétence tout en gardant l'objectif de pouvoir retracer l'évolution de ces profils de compétence dans le temps. Le souci de précision et de comparabilité des mesures ne se situe donc pas au niveau de l'individu et du diagnostic individuel, mais il se situe plutôt au niveau de la population et donc au niveau de la précision et de la validité des paramètres statistiques qui sont calculés au niveau agrégé de la population. S'ajoute à cela l'objectif d'évaluer un profil de

compétences assez large au niveau de la population, puisque ces études à large échelle sont censés évaluer les résultats du fonctionnement des systèmes scolaires vus dans leur globalité. Cela implique la nécessité d'évaluer les compétences visées sur la base de cadres de référence qui tiennent compte des différentes facettes des compétences à mesurer, ce qui implique un éventail assez large de contenus et de processus cognitifs qui doivent être couverts par ces évaluations. Cette dernière considération se heurte alors à la contrainte d'un temps de testing limité qui est disponible par sujet (en règle générale, les épreuves cognitives dans les enquêtes à large échelle ont une durée d'environ deux heures). En plus, les enquêtes à large échelle se basent sur des échantillons, mais sont censés représenter les profils de compétences des populations scolaires respectives des différents pays. Afin de trouver des solutions méthodologiques pour ces différentes contraintes, les enquêtes à large échelle utilisent en général des plans de testing incomplets sous la forme de plans d'échantillonnage multi-matrices (*multi-matrix-sampling-design*). Cela signifie que le nombre total d'items utilisés dans ces enquêtes à large échelle est nettement plus important que le nombre d'items qui peuvent être administrés à un seul individu. Concrètement, on définit des grappes d'items qui sont disjoints. Ces grappes d'items sont alors combinés de manière systématique afin de former des livrets différents qui contiennent d'un côté des grappes d'items différents, mais qui de l'autre côté présentent des intersections systématiques d'items qui vont permettre en fin de compte de réaliser un positionnement de tous les individus sur les mêmes dimensions de compétence (Smits & Vorst, 2007). Cela implique donc que différents sujets vont avoir des livrets différents contenant en partie des items différents et présentant donc un extrait limité du profil de compétences qui doit être mesuré. Au niveau de la population, l'objectif est pourtant d'arriver à une estimation des

paramètres statistiques d'un profil de compétences qui se réfère à l'ensemble des items de l'étude qui se trouvent donc distribués à travers les différents livrets.

D'après ce qui a été dit plus haut concernant les caractéristiques des MRI, il est évident qu'il sera possible de réaliser une estimation du niveau de compétences (i.e., le score thêta) de chacun des sujets sur la base d'un échantillon d'items réduit, à condition que les paramètres d'items aient été établis de manière à assurer que tous les items soient positionnés sur une même dimension de compétence. Un plan de testing avec des intersections systématiques entre des livrets différents fournit des items d'ancrage qui permettent un étalonnage de l'ensemble des items sur une même dimension. On pourrait donc penser qu'une fois les paramètres d'items établis, il suffit de réaliser pour chacun des sujets une estimation de son niveau de compétences selon les méthodes d'estimation connues pour la détermination des paramètres de sujet et d'agréger ensuite ces estimations individuelles des paramètres de sujet, afin d'établir les estimations pour les paramètres de population recherchés. Cette procédure va néanmoins impliquer des biais pour l'estimation des paramètres de population. On peut en effet montrer que la variance réelle des paramètres de sujets dans la population est surestimée dans le cas de l'utilisation d'une procédure d'estimation par individu basée sur le maximum de vraisemblance (*maximum likelihood estimate*, MLE) et que la variance est sous-estimée dans le cas de la mise en œuvre d'un algorithme d'estimation « *expected a-posteriori* » (EAP) (Mislevy, Beaton, Kaplan, & Sheehan, 1992). La mise en œuvre d'une estimation selon le maximum de vraisemblance pondérée (*weighted likelihood estimate*, WLE) implique également une sur-estimation de la variance de la population (Wu, 2005). Un problème qui est notamment associé à ces estimations biaisées des paramètres de population basés sur une simple agrégation des paramètres de sujets est le fait que cette agrégation ne tient pas compte d'une manière adéquate

de l'erreur de mesure qui est associée à chaque estimation d'un paramètre de sujet. Afin d'arriver à une estimation sans biais des différents paramètres de population qui reflète en même temps un profil de compétences basé sur l'ensemble des items qui ont été présents dans les différents livrets, il semble plus judicieux de considérer le plan incomplet qui a été obtenu à travers le plan d'échantillonnage multi-matrices comme une situation de données manquantes complètement aléatoire (missing completely at random, MCAR) (Graham, Hofer, & MacKinnon, 1996). Ainsi la problématique décrite revient à se poser la question comment on peut d'une manière adéquate tenir compte des données manquantes que chaque sujet présente par le fait qu'on lui a, d'une manière systématique, seulement présenté un nombre limité d'items, alors qu'on veut estimer dans la population les différentes statistiques pour le profil de compétences qui est représenté par l'ensemble des items présents dans l'étude. L'approche méthodologique pour tenir compte de ceci renvoie donc à la question de la gestion des données manquantes qui ont été générées d'une manière systématique et délibérée par le plan d'échantillonnage multi-matrices. Or il s'est avéré que la meilleure procédure pour tenir compte de ces données manquantes est celle des imputations multiples (Schafer & Graham, 2002). Avec la procédure des imputations multiples qui est couramment utilisée dans les études à large échelle, on va remplacer les estimations d'un paramètre de compétence unique généré par les algorithmes connus dans le contexte des MRI (MLE, WLE, EAP) par la génération de plusieurs valeurs plausibles pour la compétence des sujets qui seront autant de valeurs aléatoires tirées de la distribution postérieure du niveau de compétence d'un sujet (i.e., son score thêta) si on connaît son patron de réponses aux items qu'il a passés (Wu, 2005). Le désavantage majeur des estimations classiques est le fait qu'ils génèrent le même score pour des patrons de réponse identiques, ce qui implique une distribution discrète, alors que celle-ci est censée être continue. La méthode des valeurs plausibles qui sont donc

autant de tirages dans une distribution probabiliste va par contre générer une véritable distribution continue, car le caractère aléatoire va faire en sorte que les valeurs plausibles ne seront pas exactement identiques pour différents sujets présentant le même patron de réponse. Contrairement à l'estimation classique des paramètres de personne, les valeurs plausibles vont également fournir une information concernant l'erreur de la mesure réalisée. En effet, lorsque l'erreur de mesure sera grande, les valeurs plausibles auront tendance à varier assez fortement, alors que cette variation sera réduite pour des valeurs plausibles générées sur une mesure avec une faible erreur de mesure. Des résultats de simulation montrent que les valeurs plausibles fournissent des paramètres de population non biaisés, contrairement aux paramètres de population calculées sur la base des estimateurs classiques pour les paramètres de personne (Wu, 2005). L'imputation multiple et les valeurs plausibles constituent donc un complément indispensable à l'utilisation des MRI dans le contexte des enquêtes à large échelle. Les valeurs plausibles ne sont par contre pas adaptées à la mise en œuvre dans le contexte d'une situation de diagnostic individuel, car leur caractère de tirage aléatoire dans une distribution probabiliste avec la conséquence d'attribuer des valeurs différentes à des sujets ayant montré un même patron de réponses, n'est évidemment pas compatible avec une évaluation équitable au niveau individuel.

Un autre aspect des MRI qui est particulièrement utile dans le contexte des études à grande échelle est la possibilité de détecter un fonctionnement différentiel des items (DIF - Holland & Wainer, 1993). On parle d'un tel fonctionnement différentiel des items si les paramètres d'items diffèrent entre des populations différentes, en général suite à une familiarité plus ou moins grande de certaines populations avec le contenu de l'item. La possibilité de vérifier la présence de tels effets DIF est en effet particulièrement importante pour des études comparatives internationales, afin d'éviter que les différences constatées entre les pays ne soient

pas dus à des degrés de familiarité différents avec certains contenus qui seraient alors la suite directe des différences culturelles entre les pays et non la suite des différences dans les niveaux de compétence moyens.

Enfin signalons la grande importance des MRI dans le contexte d'analyses de tendance basées sur des items d'ancrage. L'utilisation d'items d'ancrage pour des passations successives d'études internationales telles que le PISA permet en effet de réaliser des comparaisons directes entre les vagues successives d'une même étude internationale et de calculer des indicateurs de tendance.

Testing adaptatif par ordinateur (TAO)

Un autre domaine d'application majeur des MRI est celui du testing adaptatif par ordinateur (TAO). Contrairement aux études à large échelle, le but principal d'une mise en œuvre d'un test adaptatif est plutôt la réalisation d'un diagnostic individuel d'une manière efficace tout en gardant une faible erreur de mesure. Comme pour tout test adaptatif, une caractéristique essentielle du TAO consiste en ce que les degrés de difficulté des items présentés à un sujet particulier sont choisis de manière à être adaptés au niveau de compétence de ce sujet. Un autre élément adaptatif du TAO est le fait qu'on va adapter la longueur du test de manière à faire passer seulement le nombre d'items nécessaires pour atteindre un objectif prédéterminé de qualité de mesure (qui s'exprime en général à travers l'erreur de mesure). Le but est donc d'arriver à une individualisation de la passation d'un test qui permette d'utiliser exclusivement les items les mieux adaptés pour déterminer efficacement et précisément le niveau de compétence du sujet (Dechef & Laveault, 1999; Martin, 2003; Wainer, et al., 2000). Or, de même que pour les plans d'échantillonnage multi-matrices, ce caractère adaptatif a pour conséquence que des sujets différents ne vont pas passer les mêmes items, ce qui pose encore

une fois le problème de la comparabilité des résultats obtenus. En effet, le score brut qui est utilisé dans la théorie classique du score vrai n'est plus utilisable dans le cadre du testing adaptatif, puisque le choix adaptatif des items va normalement impliquer que tous les sujets vont réussir environ la moitié des items. La mise en œuvre efficace du testing adaptatif nécessite donc un modèle de mesure qui permette de situer les sujets sur une même dimension de compétence sur la base de résultats obtenus sur des échantillons d'items différents. C'est là encore que l'utilisation des MRI s'avère indispensable. Les algorithmes les plus couramment employés pour l'estimation de la compétence ont déjà été mentionnés plus haut, à savoir les algorithmes MLE, WLE et EAP. Il reste encore à clarifier la gestion du caractère adaptatif dans le TAO, c'est-à-dire le choix des items et le critère de fin d'examen. Là encore, on fait le plus souvent appel à des éléments qui sont directement reliés aux MRI. Pour le choix des items, on adopte le plus souvent la règle du maximum d'information qui consiste à choisir comme item suivant celui qui fournit le maximum d'information au niveau de compétence estimée actuelle du sujet (le concept d'information étant défini dans le cadre des MRI). Pour déterminer la fin de la séance d'examen, on se donne en général un critère d'arrêt relatif au degré de précision visé par la mesure. Ce critère peut facilement être quantifié dans le cadre des MRI qui permettent de déterminer l'erreur de mesure associée à une estimation de compétence donnée. On va donc arrêter le test si la précision voulue est atteinte ou si la banque d'items ne fournit plus d'items susceptibles de faire diminuer l'erreur de mesure.

Dès la mise au point théorique du TAO, les avantages de cette forme d'informatisation des tests semblaient être évidents. Non seulement, on s'attendait à une individualisation du testing allant de pair avec une efficacité accrue de la passation (une meilleure précision de la mesure avec moins d'items), mais on voyait encore d'autres avantages dans la passation

informatisée, comme une sécurité plus élevée du test, une évaluation automatique et immédiate et la possibilité de développer de nouveaux formats d'items (Green, 1983; Martin, 2003). Ces perspectives alléchantes ont eu pour effet qu'au courant des années 90, on a vu notamment aux États-Unis qu'un certain nombre de programmes d'évaluation à grande échelle qui fonctionnaient depuis longtemps avec des tests sous format papier-crayon, ont été transposés à un format TAO.

Un problème majeur qui est néanmoins apparu et qui se pose notamment pour des évaluations certificatives présentant un enjeu important pour les sujets, est celui de la sécurité des épreuves. Il s'est en effet avéré que suite à l'algorithme de choix des items (le plus souvent l'algorithme du maximum d'information), on constate que la probabilité qu'un item spécifique soit choisi lors d'une passation donnée varie considérablement en fonction de la difficulté des items. Ainsi certains items sont très souvent sélectionnés au point qu'entre 15 et 20 % des items constituant une banque d'items représentent plus de 50 % des items qui sont administrés lors des passations réelles, alors que d'autres items ne sont que très rarement administrés (Wainer, 2000). Ceci a pour conséquence que le «vol d'items» par le biais d'une mémorisation systématique d'un certain nombre d'items par des sujets passant successivement le même test peut sérieusement compromettre la sécurité de l'épreuve.

Avenues de recherches futures

En guise de conclusion, nous aimerions aborder les avenues de recherches futures. Elles sont au nombre de cinq. A côté des développements directement liés aux MRI (étalonnage et estimation des paramètres de sujet), les développements méthodologiques liés aux algorithmes de sélection adaptatif des items et au contrôle de leur fréquence de présentation constituent donc un domaine de recherche qui aura une influence directe sur le succès et la faisabilité du testing

adaptatif par ordinateur (Barrada, Olea, & Ponsoda, 2007; van der Linden, Glas, Rao, & Sinharay, 2006).

Un autre aspect qui doit encore être développé à l'avenir dans le domaine des tests assistés par ordinateur d'une manière générale est celui d'une plus ample exploitation des possibilités d'affichage et de retour qui sont offertes par l'ordinateur. Ce sont avant tout ces développements qui risquent d'avoir un impact sur le développement de nouveaux modèles de mesure qui doivent être mis en œuvre afin de rendre ces items innovants pleinement opérationnels. Les tests assistés par ordinateur qui ont été mis en pratique jusqu'à aujourd'hui recourent avant tout à des types d'items dont la passation serait également possible sous forme papier-crayon, ce qui signifie qu'on peut également recourir aux mêmes modèles de mesure, notamment aux MRI. Or des travaux en sciences cognitives (et notamment en neuropsychologie cognitive, cf. par exemple Engle, Kane, & Tuholski, 1999) montrent que les performances dans des tâches complexes de type résolution de problèmes telles qu'elles sont classiquement utilisées dans les évaluations scolaires, reposent sur des processus cognitifs plus élémentaires de type perceptif ou mnésique qui sont souvent difficilement mesurables avec des dispositifs papier-crayon. Par exemple, des dispositifs d'évaluation de la mémoire de travail impliquent souvent des éléments dynamiques qui pourraient tirer pleinement profit des possibilités d'affichage dynamique qui sont offertes par l'ordinateur. Les possibilités d'affichage enrichies de l'ordinateur (et notamment sa potentialité multimédia) permettent également l'évaluation d'aptitudes qui étaient jusqu'ici largement négligées par la psychométrie, suite à la difficulté de mise en oeuvre des dispositifs d'évaluation y afférents. Vispoel (1999) présente, par exemple, un test sur ordinateur pour l'évaluation de l'aptitude musicale.

Un aspect technique qui est offert par l'ordinateur notamment dans le contexte de telles épreuves dynamiques et qui reste encore largement sous-exploité est l'utilisation des temps de réaction enregistrés lors de la résolution des items. Or comme cela a été souligné par Martin et Houssemand (2002), la signification exacte de ces temps de réaction reste encore largement indéterminée et nécessite encore des efforts théoriques majeurs. Si on veut tenir compte de ces temps de réponse d'une manière explicite dans les modèles de mesure qui sont actuellement utilisés, il faudra en plus étendre ces modèles de mesure afin d'intégrer des paramètres modélisant les temps de réponse des sujets au-delà des paramètres qui décrivent le comportement de réponse des sujets aux items. On a ainsi essayé avec succès d'étendre les MRI dans le sens qui vient d'être décrit (Ferrando & Lorenzo-Seva, 2007; Thissen, 1983).

Un autre type de tâche qui est encore assez peu exploré quant à son utilité psychométrique est l'utilisation de simulations qu'il est possible de réaliser à l'aide de l'ordinateur et qui pourraient donner accès à des tâches écologiquement plus valides pour l'évaluation d'un grand nombre de compétences, comme par exemple certaines compétences professionnelles. Ainsi Hanson, Bormann, Mogilka, Manning et Hedge (1999) présentent un dispositif de simulation pour contrôleurs de trafic aérien sur la base duquel ils génèrent des questions à choix multiple permettant d'évaluer l'aptitude des sujets à réagir d'une manière adéquate à des situations-problème qui se présentent dans ce domaine professionnel précis. Mais aussi l'évaluation de compétences plus directement liées à un certain type de fonctionnement cognitif, comme par exemple l'évaluation de la cognition spatiale, pourraient tirer bénéfice de l'utilisation de simulations informatisées, comme cela a, par exemple, été montré par Martin (1999) pour des déplacements à l'intérieur d'un espace virtuel tri-dimensionnel. La valeur ajoutée la plus importante du testing assisté par ordinateur pourrait donc être vu dans la mise au

point d'environnements de résolution de problèmes complexes qui sont proches des situations de résolution de problèmes qu'on trouve dans un environnement réel (Martin, 2008). Or ces environnements de résolution de problèmes complexes ont tous en commun qu'ils génèrent des réponses complexes, puisque le sujet va évoluer dans un environnement qui lui donne une grande liberté dans le choix et le séquencement de ses actions ce qui génère des vecteurs de données avec une très haute variabilité. L'évaluation de la qualité de la démarche de résolution du sujet devient ainsi une tâche très complexe qui nécessite la prise en compte et l'intégration des différentes facettes comportementales du sujet qui ne peuvent en règle générale pas être considérées comme des facettes indépendantes (Williamson, Bejar, & Mislevy, 2006). Cela signifie qu'on se retrouve dans une situation qui nécessite l'évaluation de la qualité de la réponse donnée par le sujet sur la base d'un vecteur de données comportementales qui est très variable et qui présente un ensemble important d'éléments interdépendants. Le défi majeur qui se pose dans ce contexte consiste à inférer, à partir de ces données complexes, l'état représentationnel du sujet ainsi que ses stratégies de résolution de problèmes et de combiner ces évaluations avec un jugement qualitatif par rapport à l'état mental actuel et par rapport aux stratégies de résolution qui sont mis en œuvre. Il est évident qu'on vise ici une évaluation qui dépasse un simple jugement en terme de réponse correcte/incorrecte et qui dépasse même un jugement sous forme de crédit partiel. La mise en œuvre des MRI qui repose sur la conception d'avoir des items à réponse correcte/incorrecte ou à crédits partiels permettant ensuite un placement du sujet sur une dimension latente va donc dans ce contexte des environnements de résolution de problèmes complexes arriver à ses limites. Il faut plutôt concevoir dans ce contexte la possibilité d'arriver à une description qualitative de ces états mentaux et de ces stratégies qui permette d'un côté un classement en vue de différencier des états mentaux plus ou moins évolués et des stratégies plus

ou moins complexes, sans pour autant exclure la possibilité de mettre en évidence des processus ou des états mentaux « vicariants » (Reuchlin, 1978), c'est-à-dire qualitativement différents, mais reflétant potentiellement des mêmes niveaux de complexité ou d'efficacité.

Les MRI sous leur forme classique modélisant une relation entre la probabilité de réussite à un item et le niveau de compétence du sujet sur un trait latent ne semblent en tout cas pas fournir une réponse adéquate en vue d'une identification des ces états mentaux et de ces stratégies de traitement à partir des données complexes recueillies à l'aide de ces dispositifs. Puisqu'il sera en plus difficilement concevable que l'ensemble de ces données complexes puissent être exploitées d'une manière efficace par un évaluateur humain, il sera absolument nécessaire de développer des algorithmes adéquats permettant d'inférer les états mentaux et les processus de traitement mentionnés sur la base d'un traitement automatique de la base des données comportementales dont on dispose à l'issue de la tâche de résolution de problème complexe sur ordinateur. Le domaine du scoring automatique de tâches complexes va donc certainement constituer un domaine de recherche privilégié pour la mise au point de nouveaux modèles de mesure qui sont nécessaires pour une utilisation efficace des nouveaux dispositifs de mesure qui seront basés sur ordinateur et qui vont essayer de tirer plus pleinement profit des possibilités d'affichage enrichi qui sont offertes par les nouvelles technologies de l'information et de la communication (Williamson, Mislevy, & Bejar, 2006). Les approches méthodologiques qui sont actuellement mises en œuvre dans ce domaine essaient d'ailleurs d'exploiter des techniques de classification et d'extraction de régularités pouvant tenir compte de données complexes et de relations non-linéaires, telles que les réseaux bayesiens (Williamson, Almond, Mislevy, & Levy, 2006) ou les réseaux de neurones artificiels (Stevens & Casillas, 2006).

Comme nous venons de le voir, les avenues futures de la recherche sur les les modèles de mesure et plus particulièrement sur les MRI sont multiples. Il apparaît évident que la revue *Mesure et évaluation en éducation* doit être plus active sinon dans le développement théorique et mathématique des MRI – qui relèvent davantage des experts en statistiques – du moins dans l'application de ces modèles au monde de l'éducation. Depuis 30 ans, la publication d'articles empiriques sur les MRI est insuffisant compte tenu de potentiel évident de ces modèles pour valider des outils d'évaluation et estimer les processus cognitifs et affectifs sous-jacents aux patrons de réponse individuels.

Références

Auger, R. (1992). Une stratégie de testing adaptatif de maîtrise. *Mesure et évaluation en éducation, 15(3),* 25-32.

Barrada, J. R., Olea, J., & Ponsoda, V. (2007). Methods for restricting maximum exposure rate in computerized adaptative testing. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 3(1),* 14-23.

Bercier-Lariviere, M. et Forgette-Giroux, R. (1999). L'évaluation des apprentissages scolaires: une question de justesse. *Revue canadienne de l'éducation, 24(1)*, 169-182.

Bertrand, R. (2001). Détection des biais d'items et de personnes en testing adaptatif. *Mesure et évaluation en éducation, 24(2-3),* 1-22.

Blais, J.-G., & Ajar, D. (1991). Théorie des réponses aux items et modélisation. *Mesure et évaluation en éducation, 14(4),* 5-18.

Blais, J.-G., & Laurier, M. (1997). La détermination de l'unidimensionalité de l'ensemble des scores à un test. *Mesure et évaluation en éducation, 20(1),* 65-90.

Burton, R. (2004). Influence des distributions du trait latent et de la difficulté des items sur les estimations du modèle de Birnbaum : une étude du type Monte-Carlo. *Mesure et évaluation en éducation, 27(3),* 41-62.

Cardinet, J. (2003). Cinq dispositifs pour vérifier le progrès. *Mesure et évaluation en éducation, 26(1-2),* 51-59.

Dechef, H, & Laveault, D. (1993). Étude du fonctionnement différentiel des items à l'aide des méthodes du khi-carré, de Mantel-Haenszel et logit. *Mesure et évaluation en éducation, 16(1-2),* 5-28.

Dechef, H., & Laveault, D. (1999). Le testing adaptatif par ordinateur. *Psychologie et Psychométrie, 20(2-3)*, 151-179.

Engle, R. W., Kane, M. J., & Tuholski, S. W. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 102-134). Cambridge, UK: Cambridge University Press.

Ferrando, P. J., & Lorenzo-Seva, U. (2007). An item response theory model for incorporating response time data in binary personality items. *Applied Psychological Measurement, 31(6)*, 525-543.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31*, 197−218.

Green, B. F. (1983). The promise of tailored tests. In H. Wainer & S. Messick (Eds.), *Principals of modern psychological measurement* (pp. 69-80). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hanson, M. A., Bormann, W. C., Mogilka, H. J., Manning, C., & Hedge, J. W. (1999). Computerized assessment of skill for a highly technical job. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 197-220). Mahwah, NJ: Lawrence Erlbaum Associates.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Lafontaine, D., & Simon, M. (2008). Évaluation des systèmes éducatif. *Mesure et évaluation en éducation, 31(3)*, ??-??.

Laurier, M. (1996). Pour un diagnostic informatisé en révision de texte. *Mesure et évaluation en éducation, 18(3),* 85-106.

Loye, N. (2005). Quelques ouveaux modèles de mesure. *Mesure et évaluation en éducation, 28(3),* 51-68.

Martin, R. (1999). *Encodage spatial et intelligence*. Lille: Presses Universitaires du Septentrion.

Martin, R. (2003). Le testing adaptatif par ordinateur dans la mesure en éducation: potentialités et limites. *Psychologie et Psychométrie, 24(2-3)*, 89-116.

Martin, R. (2008). New possibilities and challenges for assessment through the use of technology. In F. Scheuermann & A. Guimarães Pereira (Eds.), *Towards a Research Agenda on Computer-Based Assessment: Challenges and Needs for European Educational Measurement* (pp. 6-9). Luxembourg: Office for Official Publications of the European Communities.

Martin, R., & Houssemand, C. (2002). Intérêts et limites de la chronométrie mentale dans la mesure psychologique. *Bulletin de Psychologie, 55(6)*, 605-614.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29(2)*, 133–161.

Raîche, G., Langevin, L., Riopel, M., & Mauffette, Y. (2006). Étude exploratoire de la dimensionalité et des facteurs expliqués par une traduction française de l'Inventaire des approches d'enseignement de Trigwell et Prosser dans trois universités québécoises. *Mesure et évaluation en éducation, 29(2),* 41-61.

Reuchlin, M. (1978). Processus vicariants et differences individuelles. *Journal de Psychologie, 2*, 133-145.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7(2)*, 147-177.

Smits, N., & Vorst, H. C. M. (2007). Reducing the length of questionnaires through structurally incomplete designs: An illustration. *Learning and Individual Differences, 17(1)*, 25-34.

Stevens, R., H., & Casillas, A. (2006). Artificial neural networks. In D. M. Williamson, R. J. Mislevy & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 259-312). Mahwah, N.J.: Lawrence Erlbaum Associates.

St-Onge, C., Valois, P., Abdous, B., & Germain, S. (in press). A Monte Carlo study of the effect of ICC estimation on the accuracy of three person-fit statistics. *Applied Psychological Measurement.*

The Joint Committee on Standards for Educational Evaluation. (2003). The student evaluation standards. How to improve evaluations of students. Thousand Oaks, CA: Corwin Press.

Thissen, D. (1983). Timed testing: An approach using item response theory. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 179-203). New York: Academic Press.

van der Linden, W. J., & Glas, C. A. W. (2006). 25 Statistical aspects of adaptive testing.In: C.R. Rao & S. Sinharay (Eds.). *Handbook of statistics* (Vol. 26, pp. 801-838): Elsevier.

Vispoel, W. P. (1999). Creating computerized adaptive tests of music aptitude: Problems, solutions and future directions. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in Computerized Assessment* (pp. 151-176). Mahwah, NJ: Lawrence Erlbaum Associates.

Wainer, H. (2000). CATs: Whither and whence. *Psicologica, 21(1-2)*, 121-133.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (Eds.). (2000). *Computerized adaptive testing: A primer (2nd edition)*. Mahwah, NJ: Lawrence Erlbaum Associates.

Williamson, D. M., Almond, R. G., Mislevy, R. J., & Levy, R. (2006). An application of bayesian networks in automated scoring of computerized simulation tasks. In D. M. Williamson, R. J. Mislevy & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 201-258). Mahwah, N.J.: Lawrence Erlbaum Associates.

Williamson, D. M., Bejar, I. I., & Mislevy, R. J. (2006). Automated scoring of complex tasks in computer-based testing: An introduction. In D. M. Williamson, R. J. Mislevy & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 1-13). Mahwah, N.J.: Lawrence Erlbaum Associates.

Williamson, D. M., Mislevy, R. J., & Bejar, I. I. (2006). *Automated scoring of complex tasks in computer-based testing*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies In Educational Evaluation, 31(2-3)*, 114-128.