Mesure et évaluation en éducation

2010, odyssée des modèles de classification diagnostique (MCD)

Nathalie Loye

Volume 33, numéro 3, 2010

Date de réception : 13 avril 2010
Date de réception de la version finale : 28 janvier 2011
Date d'acceptation : 4 février 2011

URI : https://id.erudit.org/iderudit/1024892ar
DOI : https://doi.org/10.7202/1024892ar

Aller au sommaire du numéro

Éditeur(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (imprimé)
2368-2000 (numérique)

Découvrir la revue

Citer cet article

Loye, N. (2010). 2010, odyssée des modèles de classification diagnostique (MCD). *Mesure et évaluation en éducation*, *33*(3), 75–98. https://doi.org/10.7202/1024892ar

Résumé de l'article

Cet article vise à définir les modèles de classification diagnostique (MCD) et à déterminer leur place relativement à d'autres modélisations existantes comme la TRI. Les modèles RSM, DINA et NC-RUM sont exposés plus en détail. Pour terminer, une analyse critique débouche sur des pistes de recherches théoriques et empiriques.

2010, odyssée des modèles de classification diagnostique (MCD)

Nathalie Loye
Université de Montréal

MOTS CLÉS : Modèles de classification diagnostique, diagnostic, modèles à classes latentes

Cet article vise à définir les modèles de classification diagnostique (MCD) et à déterminer leur place relativement à d'autres modélisations existantes comme la TRI. Les modèles RSM, DINA et NC-RUM sont exposés plus en détail. Pour terminer, une analyse critique débouche sur des pistes de recherches théoriques et empiriques.

KEY WORDS: Diagnostic classification models, diagnostic, latent class models

This paper gives a definition of diagnostic classification models (DCM) and purport to compare these models to models that may be more familiar like IRT. The RSM, DINA and NC-RUM models are examined in more depth. This paper is including a critical analysis and many theoretical and empirical research avenues.

PALAVRAS-CHAVE: Modelos de classificação diagnóstica, disgnóstico, modelo de classes latentes

Este artigo pretende definir os modelos de classificação diagnóstica (MCD) e determinar o seu lugar relativamente a outras modelizações existentes, como é o caso da TRI. Os modelos RSM, DINA e NC-RUM são apresentados mais em detalhe. O artigo termina com uma análise crítica que aponta pistas para investigações teóricas e empíricas.

Note de l'auteure – Toute correspondance peut être adressée comme suit : Nathalie Loye, Université de Montréal, Faculté des sciences de l'éducation. Département d'administration et fondements de l'éducation, C.P. 6128, Succursale Centre-Ville Montréal, QC, H3C 3J7, Canada, téléphone : (514) 343-2129, télécopieur : (514) 343-2497, ou par courriel à l'adresse suivante : [nathalie.loye@umontreal.ca].

Introduction

Les modèles de diagnostic cognitif sont des modèles de mesure des habiletés sous-jacentes au processus de réponse aux items d'un test (Loye, 2005). Ils s'appuient sur deux postulats :

- la probabilité de répondre correctement à un item augmente avec la maîtrise des attributs ou habiletés qui lui sont reliés, et,
- il est possible de dresser une liste d'attributs ou d'habiletés en lien avec le test.

La place accordée à ces modèles de diagnostic cognitif peut, notamment, se mesurer par le nombre de numéros de revues qui leur ont été consacrés, en tout ou en partie, durant les trois dernières années. En effet, ce sujet a fait l'objet du numéro spécial à l'hiver 2007 du *Journal of Educational Measurement* (Almond, 2007 ; Bolt, 2007 ; DiBello & Stout, 2007 ; Gierl, 2007 ; Henson, Templin & Douglas, 2007 ; Roussos, Templin & Henson, 2007 ; Stout, 2007). Peu de temps après, un article synthèse de Rupp et Templin (2008) est publié dans le dernier numéro de l'année 2008 de la revue *Measurement : Interdisciplinary Research & Perspective*, le reste de ce numéro étant constitué par trois articles commentaires qui lui sont reliés (Gierl, 2008 ; Karelitz, 2008 ; Leighton, 2008). Les analyses et commentaires sur le texte de Rupp et Templin ont ensuite majoritairement rempli le numéro suivant car ils constituent 11 des 16 textes présentés dans le premier numéro de l'année 2009 (Frey & Carstensen, 2009 ; Gorin, 2009 ; Hancock, 2009 ; Henson, 2009 ; Jiao, 2009 ; Levy, 2009 ; Maris & Bechger, 2009 ; Sinharay & Haberman, 2009 ; C. Tatsuoka, 2009 ; von Davier, 2009 ; Wilhelm & Robitzsch, 2009). Cet ensemble de réactions montre à lui seul l'intérêt que suscitent ces modèles et offre une occasion intéressante d'en réaliser une synthèse.

Nichols et ses collaborateurs avaient donné le coup d'envoi en 1995 (Nichols, Chipman & Brennan, 1995). Dans les trois dernières années, trois nouveaux ouvrages consacrés aux modèles de diagnostic cognitif ont vu le jour (Leighton & Gierl, 2007 ; Rupp, Templin & Henson, 2010 ; K. Tatsuoka, 2009). Enfin, le manuel de statistiques de Rao et Sinharay (2007) contient un double chapitre consacré au diagnostic cognitif ; le premier (DiBello, Roussos & Stout, 2007) propose une revue des différents modèles existants, alors que le second (Haberman & von Davier, 2007) est une réflexion critique.

Basé sur cette abondante documentation, le présent article cherche à mettre en évidence ce qui caractérise ces modèles, et en quoi ils s'apparentent à d'autres modélisations existantes ou s'en éloignent. Trois modèles, exposés avec un peu plus de détails, permettent de fournir des illustrations. Le texte vise également à procurer quelques renseignements pratiques à qui souhaiterait les utiliser. Des références pertinentes pour chacun des modèles cités sont fournies, ainsi qu'une liste des applications logicielles actuellement disponibles pour les appliquer à des données. Pour terminer, notre propre analyse critique de ces modèles prend appui sur notre expérience dans l'utilisation de ces modèles, ainsi que sur les commentaires retrouvés dans les différentes réactions au texte de Rupp et Templin (2008) et dans le chapitre critique de Haberman et von Davier (2007).

Définition des modèles de classification diagnostique (MCD)

En 2008, Rupp et Templin ont publié un texte clé. D'abord, leur article est l'aboutissement de plusieurs tentatives de classifications de ces modèles de mesure. Ensuite, ces auteurs proposent une nouvelle appellation qui, selon nous, trouve mieux sa place en éducation. En effet, en transformant *modèles de diagnostic cognitif* (MDC) (*cognitive diagnostic models*- CDM) en *modèles de classification diagnostique* (MCD) (*diagnostic classification model*-DCM), ils règlent un problème souvent soulevé ces dernières années relativement au manque de théorie cognitive sous-jacente au processus de réponse aux items d'un test en éducation.

En effet, cette nouvelle appellation ne sous-entend pas l'existence d'une théorie cognitive justifiant l'existence des traits latents (habiletés, attributs) servant à établir le diagnostic. Ainsi, ces traits latents peuvent théoriquement être reliés à tout aspect qui permet d'expliquer la performance des élèves, et dont le diagnostic présente un intérêt. En l'absence de théorie cognitive sur laquelle se baser, les modèles de classification diagnostique peuvent reposer sur l'observation ou l'expérimentation pour identifier les attributs à diagnostiquer. Même si les attributs étaient souvent identifiés de manière empirique par le passé (voir par exemple, Loye, 2008), l'appellation MDC créait souvent une certaine tension.

Nous retiendrons dans la suite de ce texte la formulation *modèles de classification diagnostique* et l'acronyme *MCD* en français (*DCM* en anglais). La définition des MCD proposée par Rupp et Templin (2008) est la suivante :

> Diagnostic classification models (DCM) are probabilistic, confirmatory multidimensional latent-variable models with a simple or complex loading structure. They are suitable for modelling observable categorical response variables and contain unobservable (i.e., latent) categorical predictor variables. The predictor variables are combined in compensatory and noncompensatory ways to generate latent classes. DCM enable multiple criterion-referenced interpretations and associated feedback for diagnostic purposes, which is typically provided at a relatively fine-grain size. This feedback can be, but does not have to be, based on a theory of response processing grounded in applied cognitive psychology. Some DCM are further able to handle complex sampling designs for items and respondents, as well as heterogeneity due to strategy use (p. 226).

Les MCD se caractérisent donc par une approche probabiliste, impliquant des variables indépendantes sous forme de classes latentes (catégories) qui permettent de prévoir des variables observées dichotomiques (réponse correcte ou incorrecte aux items du test) ou polychotomiques (par exemple un crédit partiel attribué aux réponses aux items : tous les points pour une bonne réponse, une partie des points pour une réponse en partie juste, pas de points pour une réponse fausse). Tel que précisé par Rupp et Templin (2008), cette définition met volontairement à l'écart les modèles pour lesquels les traits latents se distribuent sur une échelle de mesure continue et les MCD aboutissent à une classification dans deux ou plusieurs catégories, d'où l'utilisation du terme classe latente plutôt que variable latente.

La volonté d'établir un diagnostic des sujets selon plusieurs classes latentes implique une approche multidimensionnelle. C'est bien sûr la force de l'approche, mais c'est aussi la source des difficultés à produire des données empiriques qui aboutissent à la classification souhaitée. La force vient du potentiel d'une information sur la maîtrise ou non-maîtrise de plusieurs habiletés sous-jacentes à un même test. La difficulté vient du fait que les habiletés sont souvent fortement corrélées les unes aux autres, et donc difficilement dissociables les unes des autres en pratique.

Rupp et Templin (2008) voient dans les MCD une approche essentiellement confirmatoire de par la structure imposée des liens entre les items et les habiletés. Une matrice appelée Q (Tatsuoka, 1983) fournit la structure qui relie les habiletés à diagnostiquer et les items, structure qui peut être plus ou moins complexe. Une structure simple pourrait être observée lorsqu'un

item ne vise qu'une seule habileté, dans un tel cas les MCD sont peu utiles. La complexité vient avec la possibilité de combiner les habiletés de multiples manières. Dans la Matrice Q, dont les lignes représentent les items et les colonnes représentent les habiletés, les valeurs *0* indiquent qu'une habileté n'est pas nécessaire pour l'item et une valeur *1* indique qu'elle l'est.

Certains modèles, dits non compensatoires, supposent que l'ensemble des habiletés indiquées dans la Matrice Q sont requises pour produire une réponse correcte à un item, une force sur l'une d'elles ne peut alors pas compenser une faiblesse sur une autre. C'est le cas lorsque, par exemple, les habiletés décrivent la démarche complète nécessaire pour produire la bonne réponse. D'autres MCD sont compensatoires, les habiletés reliées à l'item dans la Matrice Q ne sont alors pas toutes nécessaires pour produire une bonne réponse. L'exemple de plusieurs stratégies différentes aboutissant à une bonne réponse à l'item permet d'illustrer ce phénomène. Les classes latentes sous-jacentes aux différents MCD intègrent l'une ou l'autre des deux approches, et pourraient même combiner les deux.

Que l'on considère une approche compensatoire ou non compensatoire, l'application d'un MCD avec une Matrice Q obtenue *a posteriori* aboutit la plupart du temps à des problèmes de convergence dans les estimations et d'ajustement des données (voir par exemple : Haberman & von Davier, 2007 ; Loye, 2008 ; Loye et al., sous presse). Il est donc souvent plus approprié, dans le cas de structures simples, de modéliser les données avec des analyses factorielles ou encore d'appliquer des modèles multidimensionnels de la théorie de réponses aux items (TRI). Dans ce cas, l'utilisation de la TRI aboutit à des traits latents distribués sur un continuum avec une précision statistique plus grande que les classifications obtenues avec les MCD. C'est vraiment lorsque la structure est complexe, c'est-à-dire quand les traits latents se combinent de multiples manières pour permettre de répondre correctement aux items, que les MCD prennent tout leur sens.

Notion de règles de condensation (*condensation rules*)

Les règles de condensation sont les formules de base qui permettent de combiner les traits latents de manière compensatoire ou non compensatoire pour prédire les variables observées. Ensuite, à partir de ces formules générales, les paramètres spécifiques à chaque modèle sont inclus dans ses équations. Certains modèles intègrent par exemple des paramètres d'items

pour tenir compte des écarts qui peuvent exister entre la structure théorique des items dans la Matrice Q et la réalité du processus de réponse des sujets aux items du test (par exemple, un paramètre de pseudo-chance).

Nous adopterons dans la suite de ce texte les notations proposées par Rupp et Templin (2008) et qui sont regroupées dans le tableau 1.

Tableau 1
Notations

Identifications	Notations	
Sujets	Nombre total est I	Indexés par i=1,...,I
Sujets (non différenciables dans chaque classe latente)	Nombre total est C	Indexés par c=1,...,C
Items	Nombre total est J	Indexés par j=1,…,J
Habiletés (attributs, processus)	Nombre total est K	Indexés par k=1,…,K
Réponses (variable observée)	X_{ij} est la réponse du sujet i à l'item j	
Habiletés (variables latentes)	α_{ij} est le niveau de maîtrise du sujet i de l'habileté k	
Matrice Q	q_{jk} vaut 1 lorsque l'attribut k est requis par l'item j, 0 sinon	
Maîtrise par le sujet i des habiletés requises par l'item j (variable latente)	ξ_{ij} (xi)	
Maîtrise par le sujet i de l'habileté k requise par l'item j (variable latente)	ζ_{ij} (dzêta)	

Nous présentons ici les deux règles de condensation les plus courantes. L'équation 1 combine les classes latentes de manière non compensatoire dans une règle conjonctive. Il suffit alors qu'une seule des valeurs $P(\zeta_{ijk} = 1)$ soit égale à zéro pour que la probabilité que le sujet i fournisse une bonne réponse à l'item j soit nulle. Ainsi, le sujet i doit maîtriser l'ensemble des habiletés reliées à l'item j pour espérer répondre correctement à cet item.

L'équation 2 combine les traits latents de manière compensatoire dans une règle disjonctive. Dans ce cas, il suffit que l'une des valeurs $P(\zeta_{ijk} = 1)$ soit égale à 1 pour que la probabilité que le sujet i donne une bonne réponse à l'item j soit égale à 1. Ainsi, le sujet n'a besoin que de maîtriser l'une des habiletés reliées à l'item j pour espérer répondre correctement à cet item. Une seule habileté peut donc compenser les autres.

$$\mathrm{P}(\mathrm{X}_{ij}=1)=\prod_{k=1}^{K} P(\zeta_{ijk}=1) \qquad (1)$$

$$\mathrm{P}(\mathrm{X}_{ij}=1)=1-\prod_{k=1}^{K}(1-P(\zeta_{ijk}=1)) \qquad (2)$$

Taxonomie des modèles de classification diagnostique (MCD)

Rupp et Templin (2008) ont fait l'exercice de classer les modèles existants selon le type de variables observées, le type de variables latentes et selon le mode de combinaison des variables latentes pour prédire les variables observées (compensatoire ou non compensatoire). Le résultat de leur classification est l'objet du tableau 2 qui met en évidence la grande variété des modèles existants, propose une vision d'ensemble et fournit une référence pertinente pour chacun d'eux. Notons que la différence qui existe entre la liste des modèles répertoriés par Rupp et Templin (2008) et ceux répertoriés par DiBello et ses collaborateurs (2007) tient à la nature discrète des variables latentes imposée par la définition des MCD. Par exemple, Embretson propose deux modèles multidimensionnels de la TRI basés sur une Matrice Q et non compensatoires (Embretson & Reise, 2000 ; Whitely, 1980). Ces modèles sont le *multi-component latent trait model* (MLTM) et le *general component latent trait model* (GLTM). Ils ne sont pas répertoriés ici car les paramètres représentant les habiletés sont continus dans ces deux modèles.

Tableau 2
Taxonomie (Rupp & Templin, 2008)

	Variables observées (dépendantes)		Variables latentes (indépendantes)				
Noms des modèles	dichotomique	polychotomique	dichotomique	polychotomique	compensatoire	non compensatoire	**Références**
Rule space method (**RSM**)	X	X	X			X	(Tatsuoka, 1983 ; 2009)
Skill hierarchy method (**AHM**)	X	X	X			X	(Leighton, Gierl & Hunka, 2004)
Bayesian inference network (**BIN**)	X	X	X	X	X	X	(Yan, Mislevy & Almond, 2003)
Deterministic inputs, noisy 'and' gate (**DINA**)	X		X			X	(Junker & Sijtsma, 2001)
Higherorder DINA (**HO-DINA**)	X		X			X	(de la Torre & Douglas, 2004)
Multi-strategy DINA (**MS-DINA**)	X		X			X	(de la Torre & Douglas, 2005)
Loglinear Cognitive Diagnosis Model (**LCDM**)	X	X	X	X	X		(Henson, Templin & Willse, 2009)
Deterministic inputs, noisy 'or' gate (**DINO**)	X		X		X		(Templin & Henson, 2006)
Noisy inputs, deterministic 'and' gate (**NIDA**)	X		X			X	(Junker & Sijtsma, 2001)
Noisy inputs, deterministic 'or' gate (**NIDO**)	X		X		X		(Templin & Henson, 2006)
Non-compensatory Reparametrized unified model / Fusion model (**NC-RUM**)	X	X	X	X		X	(DiBello, Stout & Roussos, 1995 ; Hartz, 2002)
Compensatory Reparametrized unified model / Fusion model (**C-RUM**)	X	X	X	X	X		(Templin & Henson, 2006)
Reduced Reparametrized unified model (**RE-RUM**)	X		X			X	(Templin & Henson, 2005)
General diagnostic model (**GDM**)	X	X	X	X	X		(von Davier, 2005)
Loglinear cognitive diagnosis model (**LCDM**)	X	X	X	X	X		(Henson, Templin & Willse, 2009)
Multiple classification latent class model (**MCLCM**)	X	X	X	X	X	X	(Maris, 1999)

Note. Adapté de Rupp, A., & Templin, J. (2008). Unique characteristics of diagnostic classification models : A comprehensive review of the current state-of-the-art. *Measurement : Interdisciplinary Research & Perspective, 6*(4), p. 239.

Disponibilité des logiciels

Le tableau 3 présente les logiciels actuellement disponibles pour appliquer les MCD à des données.

Tableau 3
Logiciels (Rupp & Templin, 2008)

Logiciel	MCD	Type de logiciel et références
BUGLIB	RSM	Licence de recherche [tatsuoka@prodigy.net]
AHM	AHM	Licence de recherche [mark.gierl@ualberta.edu]
DCM	DINA, NIDA, DINO, NIDO, NC-RUM réduit, C-RUM	Freeware à utiliser avec M-Plus [jtemplin@uga.edu]
DCM dans R	DINA, DINO	Freeware à utiliser avec R (gratuit) [alexander.robitzsch@iqb.hu-berlin.de]
DINA dans Ox	DINA, HO-DINA, MS-DINA, G-DINA	Freeware à utiliser avec Ox (gratuit) [j.delatorre@rutgers.edu]
Arpeggio	NC-RUM (complet et réduit)	Commercial [www.assess.com]
LCDM	LCDM	Freeware à utiliser avec M-Plus [jtemplin@uga.edu]
MDLTM	MDLTM	Licence de recherche [mvondavier@ets.org]

Note. Adapté de Rupp, A., & Templin, J. (2008). Unique characteristics of diagnostic classification models : A comprehensive review of the current state-of-the-art. *Measurement : Interdisciplinary Research & Perspective, 6*(4), p. 250.

Trois modèles en particulier

Nous nous attardons sur trois modèles dans les paragraphes qui suivent. Pour chacun d'eux sont présentées quelques considérations théoriques, accompagnées d'une fiche technique. Pour les deux derniers modèles, des considérations pratiques issues de notre expérience liée à leur application à des données sont ajoutées.

Le premier modèle est le *Rule space* (RSM). Nous l'avons retenu pour deux raisons. La première est son caractère intuitif qui le différencie des autres modèles, à l'exception du modèle AHM qui en est issu. Le fait que le RSM est l'un des modèles les plus anciens, et qu'il a été appliqué à des données empiriques avec succès à plusieurs reprises, constitue la deuxième raison. Par exemple, Tatsuoka a utilisé son modèle avec les données du *Scholastic aptitude*

test (SAT) (mathématique) en 1993 (Birenbaum, Kelly & Tatsuoka, 1993) et avec les données de l'étude sur les Tendances de l'enquête internationale sur les mathématique et les sciences (TEIMS) (Dogan & Tatsuoka, 2008 ; Tatsuoka, Corter & Tatsuoka, 2004). Buck et ses collaborateurs ont utilisé des données en lecture et en compréhension de texte en 1997 et 1998 (Buck, Tatsuoka & Kostin, 1997 ; Buck & Tatsuoka, 1998). Yepes-Baraya (1998) a utilisé les données en sciences du *National Assessment of Educational Progress* (NAEP).

Le RSM se caractérise par l'utilisation du modèle logistique à deux paramètres (2PL) de la TRI pour créer un plan de classification cartésien dans lequel l'axe des abscisses représente la valeur estimée de l'habileté des sujets notée θ, et l'axe des ordonnées correspond à une mesure (*caution indice*), notée ζ et issue de la TRI (Tatsuoka, 1984), de l'adéquation entre les réponses de chaque sujet et ce qui est attendu (*atypicality* ou *person fit*). L'indice ζ est calculé globalement pour l'ensemble des items du test, toutefois il est également possible de calculer ζ_1, ζ_2, ζ_3, … relativement à des sous-ensembles d'items reliés à des contenus différents (par exemple algèbre, géométrie, etc.). Les coordonnées (θ, ζ) générées pour chaque sujet correspondent à un point dans le plan cartésien.

La structure du test sous forme d'une Matrice Q est attribuable à Tatsuoka (1983). Une matrice contenant K habiletés peut ensuite théoriquement aboutir à 2^K combinaisons possibles. Toutefois, la réalité est plus parcimonieuse pour Tatsuoka qui a exploité les propriétés de l'algèbre de Boole pour identifier des états de connaissances plausibles en se basant sur le fait que certaines habiletés sont préalables à d'autres (par exemple, un élève ne peut maîtriser la multiplication que s'il sait faire une addition). Les combinaisons plausibles d'habiletés qui sont identifiées permettent de générer les patrons de réponses idéaux correspondants en se basant sur la Matrice Q. Il est alors possible d'estimer les paramètres θ et ζ de ces patrons idéaux, puis de placer les points correspondants dans le plan cartésien.

Chaque sujet et chaque état idéal correspondent donc à un point dans le plan cartésien. La première étape pour réaliser le diagnostic d'un sujet consiste à calculer et à ordonner les distances entre le point de ce sujet et les points représentant les différents états idéaux. Par la suite, les L distances plus petites qu'un seuil prédéterminé sont retenues. Dans un troisième temps, le calcul des probabilités postérieures correspondant à chacun des L états idéaux retenus permet de pondérer chaque état idéal pour calculer la probabilité de maîtrise

de chacune des habiletés. En pratique, certaines considérations peuvent être utilisées pour ajuster les états idéaux finalement inclus dans le calcul des probabilités. Pour plus d'information à ce sujet, voir le chapitre 7 dans K. Tatsuoka (2009).

Le tableau 4 présente la fiche technique du modèle *Rule space* (RSM) de Tatsuoka.

Tableau 4
Fiche technique du RSM

Fiche technique	
Références	Tatsuoka, 1983, 1995, 2009
Type	Non compensatoire
Scores	Dichotomiques
Classes latentes	Dichotomiques ou polychotomiques
Paramètres spécifiques	Méthode analytique et non statistique
Équation	Le RSM n'est pas un modèle statistique donc il n'y a pas d'équation représentant le modèle
	Le RSM transforme les données en probabilités de maîtrise des habiletés.
	Il nécessite une matrice Q et des états de connaissance plausibles qui permettent de faire le lien entre observable et latent
Logiciel	BUGLIB

Note. Adapté de Rupp, A. A. (2009, avril) *Software for calibrating Diagnostic Classification Models.* Symposium conduit lors de l'American Educational Research Association de San Diego, CA. Documentation disponible à : [http ://www.education.umd.edu/EDMS/fac/Rupp/].

Le deuxième modèle a été choisi en raison de sa simplicité. Le modèle DINA (*deterministic inputs, noisy 'and' gate model*) (de la Torre & Douglas, 2004 ; Junker & Sijtsma, 2001) est un modèle non compensatoire pour lequel les données doivent être dichotomiques. Il se caractérise par deux paramètres d'item. Le paramètre de pseudo-chance g_j permet de prendre en considération le fait qu'un individu devine la réponse à l'item *j* au lieu de la trouver grâce aux habiletés identifiées dans la Matrice Q. Le paramètre d'étourderie s_j correspond au cas où un individu maîtrisant toutes les habiletés requises fournit une mauvaise réponse à un item. Ainsi, la probabilité de donner une bonne réponse à un item peut être $(1\text{-}s_j)$ ou (g_j) selon que les habiletés sont ou non maîtrisées. L'objectif est de maximiser la différence entre la probabilité de

bien répondre à l'item selon que l'on possède ($1-s_j$) ou pas (g_j) les habiletés spécifiées. La principale limite de ce modèle tient au fait que la probabilité de bonne réponse ne tient pas compte du nombre ou du type d'habiletés qui ne sont pas maîtrisées (Roussos et al., 2007). Notons que quelques applications empiriques ont été tentées, avec un succès parfois mitigé (de la Torre, 2008 ; de la Torre & Douglas, 2004 ; Loye et al., sous presse), probablement partiellement à cause de cette limite.

Le tableau 5 présente la fiche technique du modèle DINA. Les méthodes d'estimation des paramètres y sont mentionnées, ainsi que des références pertinentes. Le logiciel Ox peut être téléchargé gratuitement pour tout usage relié à la recherche. L'algorithme peut être obtenu gratuitement sur demande auprès de Jimmy de la Torre [j.delatorre@rutgers.edu]. De plus, de la Torre propose une présentation didactique du modèle DINA et de l'estimation de ses paramètres (de la Torre, 2009).

Tableau 5
Fiche technique du DINA

Fiche technique	
Références	Macready & Dayton, 1977 Junker & Sitjma. 2001 de la Torre & Douglas, 2004
Type	Non compensatoire
Scores	Dichotomiques
Classes latentes	Dichotomiques
Paramètres spécifiques	Pseudo chance (g_j)
	Étourderie (s_j)
Équation	$P(X_{ij}=1 \mid \xi_{jc}, s_j, g_j) = (1-s_j)^{\xi_{jc}} g_j^{1-\xi_{jc}}$
Estimation	MMLE (Bock & Aitkin, 1981) MCMC (de la Torre & Douglas, 2004) EM (Gitomer & Rock, 1993 ; Haertel, 1984, 1990)
Logiciel	Algorithme dans Ox (Doornik, 2002) [http ://www.oxmetrics.net/]

Une fois le logiciel Ox installé sur un ordinateur, il suffit de placer dans un même dossier le fichier contenant l'algorithme DINA et les deux fichiers textes contenant les données d'une part et la Matrice Q d'autre part. Après avoir spécifié convenablement les noms de fichiers et les nombres de sujets (I)[1], d'items (J) et d'habiletés (K) dans l'algorithme et lancé l'estimation, deux nouveaux fichiers sont produits (alpha.out et beta.out). Les deux fichiers obtenus peuvent être visualisés avec Ox. La figure 1 donne un exemple du dossier obtenu. Le fichier alpha.out fournit le vecteur diagnostique de chaque sujet sous forme d'une suite de valeurs *0* ou *1* pour les K habiletés incluses dans la Matrice Q. Le fichier beta.out contient les paramètres g_j et s_j pour chaque item et leurs écarts types.

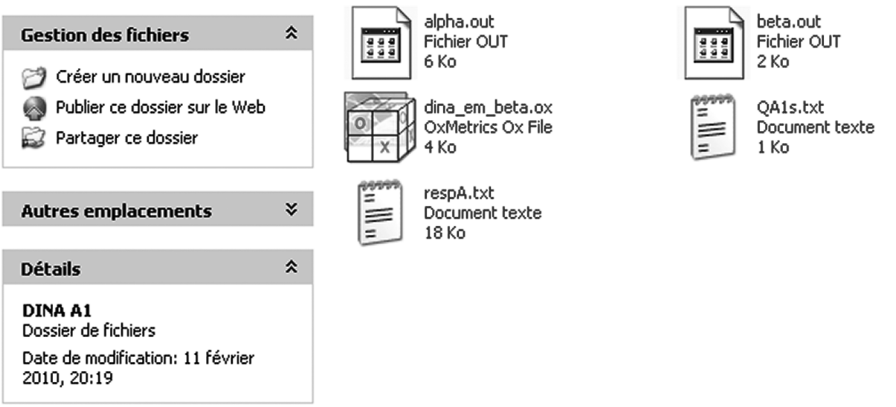

Figure 1. ***Le contenu du dossier DINA après estimation.***

Enfin, nous avons retenu le troisième modèle parce qu'il est l'un des plus complexe et qu'il est applicable à des données. Le modèle NC-RUM (Hartz, 2002) est un modèle non compensatoire pour lequel les données peuvent être de type dichotomiques ou polychotomiques. Il se caractérise par trois paramètres d'item. Le paramètre π_j représente la probabilité qu'un sujet qui maîtrise les habiletés requises par un item les utilise convenablement pour répondre à la question. Le paramètre r_j représente la pénalité due au fait de ne pas maîtriser une habileté. Enfin, le troisième paramètre c_j permet de juger si la Matrice Q contient toutes les habiletés importantes, ce dernier paramètre est fixé dans la version réduite du modèle. Ce modèle tient donc compte du nombre ou du type d'habiletés qui ne sont pas maîtrisées pour calculer la probabilité d'une bonne réponse.

De plus, les habiletés sont caractérisées par leur difficulté, notée p_k, qui correspond à la probabilité de maîtriser l'habileté k. Cette probabilité est estimée pour chaque sujet ainsi que globalement. Notons que plusieurs applications à des données sont également disponibles dans la documentation, par exemple en mathématique (Loye, 2008, 2009 ; Yan, Almond & Mislevy, 2003) et en lecture en anglais langue seconde (Jang, 2005).

Le tableau 6 présente la fiche technique du modèle NC-RUM. Les méthodes d'estimation des paramètres y sont mentionnées ainsi que des références pertinentes. Le logiciel Arpeggio est vendu sous la forme d'un disque compact dans lequel est inclus le manuel d'utilisation contenant une liste d'exercices. Tout comme dans le cas de Ox pour DINA, il convient de créer un dossier pour chaque analyse. Ce dossier doit contenir les quatre fichiers d'Arpeggio (fichiers.exe), le fichier texte contenant l'algorithme (arpeggio.in) ainsi que les deux fichiers textes contenant les données et la Matrice Q, tous les fichiers doivent avoir une extension *.in* et non pas *.txt*.

Tableau 6
Fiche technique du NC-RUM

Fiche technique	
Références	DiBello et al., 1995 Hartz, 2002 Roussos et al., 2007
Type	Non compensatoire (existe en version compensatoire)
Scores	Dichotomiques ou polychotomiques
Classes latentes	Dichotomiques ou polychotomiques
Paramètres spécifiques	π_j : difficulté de l'item j relativement aux habiletés reliées r_{jk} : pénalité due au fait de ne pas maîtriser l'attribut k c_j : exhaustivité de la liste d'attributs
Équation (version NC-RUM réduit)	$P(X_{ij}=1 \mid \pi^*_j, r^*_{jk}) = \pi^*_j \prod_{k=1}^{K} r_j^{*(1-\alpha_{ck})q_{jk}}$
Estimation	MCMC (Hartz, 2002) EM (Gitomer & Rock, 1993 ; Haertel, 1984, 1990)
Logiciel	Arpeggio [http://www.assess.com/xcart/product.php?productid=437&cat=1&page=1]

Après avoir spécifié convenablement les noms de fichiers dans l'algorithme (arpeggio.in), fait divers choix relatifs aux estimations (voir le manuel d'Arpeggio pour plus de détails), le processus d'estimation peut commencer. C'est le fichier arpeggio3_1.exe qui permet de mettre en route l'analyse ; celle-ci commence une fois que le nom du fichier contenant l'algorithme est entré dans la fenêtre. La figure 2 présente un exemple de dossier obtenu après estimation ; celui-ci contient sept fichiers initiaux et les dix fichiers produits, contenant notamment les paramètres estimés.

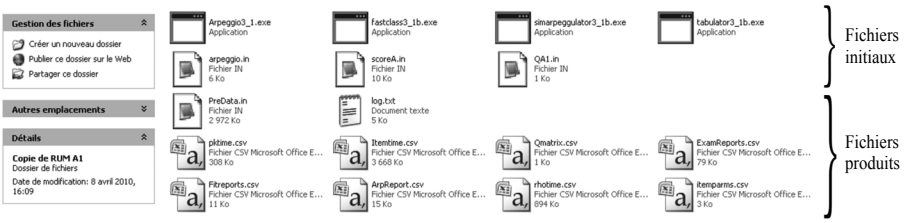

Figure 2. ***Le contenu du dossier RUM après estimation.***

Les fichiers predata.in, log.in et Qmatrix.csv sont des récapitulatifs des analyses et des données. Les fichiers pktime.csv, itemtime.csv fournissent les estimations des paramètres p_k et d'items au fil des chaînes de Markov. Il est ensuite possible de vérifier si les chaînes convergent vers une valeur avant de chercher à interpréter les paramètres obtenus. Une méthode consiste à représenter graphiquement les valeurs au fil des chaînes avec le logiciel R [http://www.r-project.org/] en utilisant un code disponible à l'adresse [http://cran.r-project.org/web/packages/coda/index.html]. Un exemple de graphique ainsi obtenu fait l'objet de la figure 3.

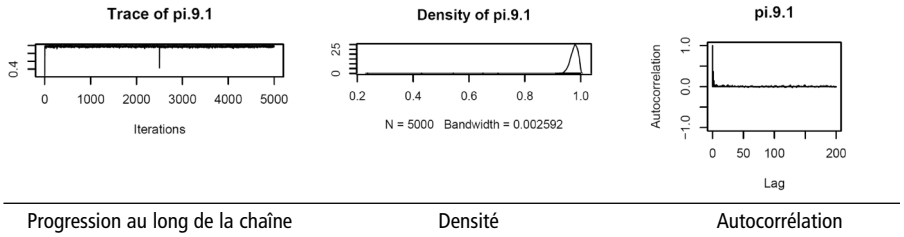

Figure 3. ***Exemple de graphiques montrant la convergence d'un paramètre π.***

Le fichier ExamReport.csv contient les probabilités de maîtrise de chaque habileté par chaque sujet alors que le fichier itemparms.csv contient les paramètres d'items estimés. Les deux derniers fichiers renferment de nombreux renseignements permettant de vérifier l'ajustement des données. Le manuel d'Arpeggio fournit les explications nécessaires à la compréhension du contenu de chaque fichier, ainsi que des exemples commentés.

Regard critique et points sensibles

Le fait que ces modèles soient à la fois complexes et récents est à l'origine de multiples regards critiques qui font l'objet des paragraphes qui suivent. Seront abordés la complexité des modèles, les problèmes liés à leur validité, la nature des habiletés, le format du diagnostic et le manque d'études empiriques.

Plusieurs auteurs questionnent la nécessité d'utiliser des modèles aussi complexes. Leurs arguments sont, d'une part, que la preuve n'a pas été faite que ces modèles apportent des données plus pertinentes que d'autres modèles plus simples, mieux connus et plus adaptés aux données disponibles en éducation qui sont souvent unidimensionnelles. Par exemple, Gorin (2009) remet en question une prémisse voulant que la multidimensionnalité soit nécessaire pour permettre un diagnostic en rappelant l'utilisation des cartes de Wright pour relier les réponses aux items d'un test unidimensionnel à des processus cognitifs sous-jacents. Rappelons qu'une carte de Wright est une représentation graphique dans laquelle la difficulté des items et l'habileté des candidats sont placées sur une même échelle de mesure. D'autre part, plusieurs de ces modèles restent théoriques car leurs paramètres ne sont pas identifiables (Maris & Bechger, 2009), ou du moins un doute persiste quant à la possibilité d'estimer les paramètres en pratique. Cela signifie que l'usage de ces modèles pour des besoins pratiques liés à la salle de classe n'est pas pour demain, d'autant que la complexité va actuellement avec des bases de données de très grande taille.

L'un des aspects les plus problématiques vient du manque d'études de validité pour ces modèles. La validité externe des classifications diagnostiques obtenues doit être étudiée en comparant les résultats à d'autres sources d'information (Bolt, 2007 ; Haberman & von Davier, 2007). L'équivalence des diagnostics issus des mêmes données mais de modèles différents, ou de données provenant de différents tests doit également être étudiée (Maris & Bechger, 2009 ; Roussos et al., 2007 ; Sinharay & Haberman, 2009). Enfin, la

validité interne des modèles doit faire l'objet d'études approfondies afin de s'assurer de la qualité des classifications (Roussos et al., 2007 ; Sinharay & Haberman, 2009). En outre, les procédures pour vérifier l'ajustement des données aux modèles sont quasi absentes (Levy, 2009). Seul Arpeggio et les modèles RUM fournissent des renseignements permettant de juger la convergence des estimations (Roussos et al., 2007).

La nature des habiletés sur lesquelles faire porter le diagnostic est également source de nombreuses discussions dans la documentation. Un grand manque d'études empiriques est d'abord à noter (Bolt, 2007 ; Sinharay & Haberman, 2009) ainsi que le fait que les études empiriques existantes consistent à appliquer un MCD à des données non initialement prévues pour cet usage. À l'heure actuelle, le développement de tests est basé sur des modèles de la TRI et l'objectif est le plus souvent de mesurer un seul construit à la fois. Demander à des experts d'identifier des habiletés sous-jacentes à ce construit et tenter de réaliser un diagnostic de ces habiletés est appelé *retrofitting* et cette approche est largement remise en question. D'abord, le fait de se baser sur le jugement d'experts pour formaliser la Matrice Q ajoute à la complexité car il faut définir quels aspects sont sous la responsabilité des experts et lesquels proviennent directement des données. En outre, le développement de tests est basé sur un paradigme différent de celui sur lequel reposent les MCD (Haberman & von Davier, 2007). Dans un cas, l'hypothèse est l'existence d'un (ou éventuellement de plusieurs) trait latent continu et dans l'autre, celle de l'existence d'un ensemble de variables latentes discrètes. Ainsi, dans un cas les différentes dimensions sont supposées être hautement corrélées pour ne faire qu'une et être interchangeables, alors que dans l'autre les habiletés peuvent se combiner différemment d'un sujet à un autre.

Analyser les habiletés peut donc être vu comme une manière d'étudier l'interaction entre le sujet et l'item. Ainsi, les habiletés sont à l'origine du choix du modèle, selon qu'elles sont vues comme compensatoires ou non, et le test devrait être développé sur la base de ces habiletés. Il faut donc définir de nouvelles méthodes pour développer des tests et vérifier la fiabilité des données obtenues avant de pouvoir espérer mener des études empiriques valides (Gorin, 2009 ; Henson, 2009 ; Loye et al., sous presse ; Roussos et al., 2007 ; Sinharay & Haberman, 2009). Toutefois, la question de savoir s'il est vraiment possible de démêler des habiletés spécifiques qui sont fortement corrélées entre elles, ainsi qu'à une habileté générale, reste ouverte. La nature discrète des habiletés est aussi questionnable (Henson, 2009 ; Levy, 2009).

Pour terminer, le format dans lequel le diagnostic devrait être proposé aux enseignants pour qu'ils en tirent profit est également source de discussions et divers exemples sont disponibles (DiBello et al., 2007). À ce propos, le format dichotomique proposé par de nombreux modèles peut sembler n'être pas optimal pour le diagnostic (Karelitz, 2008). Finalement, c'est aussi la valeur ajoutée de ce diagnostic qui est remise en question. Les enseignants voient-ils un intérêt à obtenir un rapport en fonction des habiletés de leurs élèves (Haberman & von Davier, 2007) ? Probablement que oui, étant donné la place que l'évaluation formative a prise dans les curricula. Toutefois, ce rapport diagnostique ne peut être utile que s'il est simple à lire et à comprendre et s'il est accompagné de pistes de remédiation en lien avec les difficultés ciblées.

Malgré les critiques, les MCD prennent de plus en plus de place dans la documentation et de nombreux chercheurs travaillent à développer ces modèles, à les rendre plus faciles à utiliser, à limiter la taille des bases de données nécessaires ou encore à revoir le format des rapports diagnostiques. Même si la majorité des articles concernant les MCD sont axés sur les aspects théoriques et statistiques et utilisent des données simulées, plusieurs applications empiriques avec des données réelles les complètent.

Pistes multiples de recherche

À partir des critiques et des points sensibles liés aux MCD, les pistes de recherche peuvent être classées en deux catégories. Tout d'abord, des études théoriques sont nécessaires, notamment pour fournir des modèles plus flexibles et donc mieux adaptés à la réalité. Deux objectifs peuvent être mentionnés à cet effet, comme minimiser la taille des bases de données nécessaires, ou encore combiner les approches compensatoires et non compensatoires. Plusieurs des modèles existants restent théoriques et ne sont pas faciles à utiliser par la communauté des chercheurs. Dans certains cas, ces modèles ne sont pas identifiables ; dans d'autres, aucun algorithme n'est disponible pour les appliquer à des données.

Peu de balises existent dans la documentation quant aux nombres d'items, d'habiletés ou de sujets qui doivent être considérés pour une application pratique et valide de ces modèles. Des études doivent donc être menées pour fournir de telles balises.

La validité des modèles doit être étudiée avec attention, car c'est l'une des critiques majeures de ces modèles à l'heure actuelle. Dans ce sens, des recherches doivent porter sur l'équivalence des modèles et des diagnostics issus des modélisations. Enfin, il convient aussi de doter les différents modèles de mesures pratiques permettant de s'assurer de la convergence des algorithmes, de l'ajustement des données aux modèles ou du dépistage des sujets ayant des schémas de réponses anormaux.

Enfin, les recherches empiriques doivent se multiplier. Pour ce faire, un aspect préalable important consiste à développer des tests permettant de générer des données ayant un pouvoir diagnostique en accord avec les postulats de ces modèles. Cet aspect passe peut-être par le développement de nouvelles manières de créer des items, de les combiner et d'attribuer des scores.

Conclusion

Le présent article prend sa source dans le texte publié par Rupp et Templin en 2008. Il vise à mettre en évidence ce qui caractérise les MCD et fournit de nombreuses références récentes. Les MCD y sont exposés de manière générale et trois modèles pour lesquels des applications logicielles sont disponibles sont plus particulièrement présentés. Le texte inclut une analyse critique des MCD qui débouche sur plusieurs pistes de recherche.

Le titre de cet article parle d'odyssée, définie comme un voyage riche en péripéties. Dans le cas des MCD, nous sommes au début du voyage, mais le nombre et la richesse des études actuelles laissent présager que les MCD vont continuer à se développer dans les années à venir. Ces modèles offrent donc de belles perspectives de recherche autant théoriques qu'appliquées.

NOTE

1. Selon les notations du tableau 1.

RÉFÉRENCES

Almond, R. (2007). Modeling diagnostic assessments with Bayesian networks. *Journal of Educational Measurement, 44*(4), 341-359.

Birenbaum, M., Kelly, A. E., & Tatsuoka, K. K. (1993). Diagnosing knowledge states in algebra using the rule-space model. *Journal for Research in Mathematics Education, 24*(5), 442-459.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. *Psychometrika, 46*, 443-449.

Bolt, D. (2007). The present and the future of IRT-based cognitive models (ICDMs) and related methods. *Journal of Educational Measurement, 44*(4), 377-383.

Buck, G., & Tatsuoka, K. K. (1998). Application of the rule space procedure to language testing: examining attributes of a free response listening test. *Language Testing, 15*(2), 119-157.

Buck, G., Tatsuoka, K. K., & Kostin, I. (1997). The subskills of reading: Rule-space analysis of a multiple choice test of second language reading comprehension. *Language Testing, 47*(3), 423-466.

de la Torre, J. (2008). An empirically-based method of Q-matrix validation for the DINA model : Development and applications. *Journal of Educational Measurement, 45*(4), 343-362.

de la Torre, J. (2009). DINA model and parameter estimation : A didactic. *Journal of Educational and Behavioral Statistics, 34*(1), 115-130.

de la Torre, J., & Douglas, J. A. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika, 69*(3), 333-353.

de la Torre, J., & Douglas, J. A. (2005, april). *Modeling multiple strategies in cognitive diagnosis*. Article présenté au congrès annuel du National Council on Measurement in Education (NCME), Montréal, QC.

DiBello, L. V., Roussos, L., & Stout, W. (2007). Review of cognitively diagnostic assessment and a summary of psychometric models. In C. R. Rao & S. Sinharay (dir.), *Handbook of Statistics* (vol. 26, pp. 979-1030). Amsterdam : Elsevier.

DiBello, L. V., & Stout, W. (2007). Guest editors' introduction and overview: IRT-based cognitive diagnostic models and related methods. *Journal of Educational Measurement, 44*(4), 285-291.

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likeklihood-based classification techniques. In P. D. Nichols, S. F. Chipman & R. L. Brennan (dir.), *Cognitively diagnostic assessment* (pp. 361-389). Hillsdale, NJ: Erlbaum.

Dogan, E., & Tatsuoka, K. (2008). An international comparison using a diagnostic testing model: Turkish students' profile of mathematical skills on TIMSS-R. *Educational Studies in Mathematics, 68*(3), 263-272.

Doornik, J. A. (2002). Object-oriented matrix programming using Ox (version 3.1) [Logiciel]. London : Timberlake Consultats Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, N.J. : Lawrence Erlbaum Associates.

Frey, A., & Carstensen, C. H. (2009). Diagnostic classification models and multidimensional adaptive testing : A commentary on Rupp and Templin. *Measurement: Interdisciplinary Research & Perspective, 7*(1), 58-61.

Gierl, M. (2007). Making diagnostic inferences about cognitive attributes using the Rule-Space Model and Attribute Hierarchy Method. *Journal of Educational Measurement, 44*(4), 325-340.

Gierl, M. (2008). Defining characteristics of diagnostic classification models and the problem of retrofitting in cognitive diagnostic assessment. *Measurement: Interdisciplinary Research & Perspective, 6*(4), 263-268.

Gitomer, D. H., & Rock, D. (1993). Adressing process variables in test analysis. In N. Fredericksen, R. J. Mislevy & I. I. Bejar (dir.), *Test theory for a new generation of tests* (pp. 125-150). Hillsdale, NJ : Erlbaum.

Gorin, J. S. (2009). Diagnostic classification models: Are they necessary? Commentary on Rupp and Templin (2008). *Measurement: Interdisciplinary Research & Perspective, 7*(1), 30-33.

Haberman, S. J., & von Davier, M. (2007). Some notes on models for cognitively based skills diagnosis. In C. R. Rao & S. Sinharay (dir.), *Handbook of Statistics* (vol. 26, pp. 1031-1039). Amsterdam: Elsevier.

Haertel, E. H. (1984). An application of latent class models to assessment data. *Applied Psychological Measurement, 8*, 333-346.

Haertel, E. H. (1990). Continuous and discrete latent class structure models of item response data. *Psychometrika, 55*, 477-494.

Hancock, G. R. (2009). Diagnostic classification modeling: opportunity for identity. *Measurement: Interdisciplinary Research & Perspective, 7*(1), 62-64.

Hartz, S. M. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities: Blending theory with practicality.* Dissertation doctorale non publié, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.

Henson, R. (2009). Diagnostic classification models: throughts future directions. *Measurement: Interdisciplinary Research & Perspective, 7*(1), 34-36.

Henson, R., Templin, J., & Douglas, J. (2007). Using efficient model based sum-scores for conducting skills diagnoses. *Journal of Educational Measurement, 44*(4), 361-376.

Henson, R., Templin, J., & Willse, J. (2009). Defining a family of cognitive diagnosis models using log-linear models with latent variables. *Psychometrika, 74*(2), 191-210.

Jang, E. E. (2005). *A validity narrative : Effects of reading skills diagnosis on teaching and learning in the context of NG TOEFL*. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.

Jiao, H. (2009). Diagnostic classification models: Which one should I use? *Measurement: Interdisciplinary Research & Perspective, 7*(1), 65-67.

Junker, B. W., & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25*, 258–272.

Karelitz, T. (2008). How binary skills obscure the transition from non-mastery to mastery. *Measurement: Interdisciplinary Research & Perspective, 6*(4), 268-272.

Leighton, J. (2008). Where's the psychology? A commentary on "unique characteristics of diagnostic classification models : A comprehensive review of the current state-of-the-art". *Measurement: Interdisciplinary Research & Perspective, 6*(4), 272-275.

Leighton, J. P., & Gierl, M. J. (2007). *Cognitive diagnostic assessment for education : Theory and applications*. Cambridge: Cambridge University Press.

Leighton, J. P., Gierl, M. J., & Hunka, S. M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space approach. *Journal of Educational Measurement, 41*, 205-237.

Levy, R. (2009). Evidentiary reasonning in diagnostic classification models. *Measurement: Interdisciplinary Research & Perspective, 7*(1), 36-41.

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

Loye, N. (2008). *Conditions d'élaboration de la Matrice Q des modèles cognitifs et impact sur sa validité et sa fidélité*. Thèse de doctorat non publiée, Université d'Ottawa, Ottawa.

Loye, N. (2009). Les modèles cognitifs. In J.-G. Blais (dir.), *Évaluation des apprentissages et technologies de l'information et de la communication : Enjeux, applications et modèles de mesure*. Québec : PUL.

Loye, N., Caron, F., Pineault, J., Tessier-Baillargeon, M., Burney-Vincent, C., & Gagnon, M. (sous presse). La validité du diagnostic issu d'un mariage entre didactique et mesure sur un test existant. In G. Raîche, K. Paquette-Côté & D. Magis (dir.), *Des mécanismes pour assurer la validité de l'interprétation de la mesure en éducation* (vol. 1). Sainte-Foy, Québec : Presses de l'Université du Québec.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2*, 99-120.

Maris, E. (1999). Estimating multiple classification latent class models. *Psychometrika, 64*, 187-212.

Maris, G., & Bechger, T. (2009). Equivalent diagnostic classification models. *Measurement : Interdisciplinary Research & Perspective, 7*(1), 41-46.

Nichols, P. D., Chipman, S. F., & Brennan, R. L. (1995). *Cognitively diagnostic assessment*. Hillsdale, NJ: Erlbaum.

Rao, C. R., & Sinharay, S. (dir.). (2007). *Handbook of statistics* (vol. 26). Amsterdam: Elsevier.

Roussos, L., Templin, J., & Henson, R. (2007). Skills diagnosis using IRT-based latent class models. *Journal of Educational Measurement, 44*(4), 293-311.

Rupp, A. A. (2009, avril). *Software for calibrating Diagnostic Classification Models*. Symposium conduit lors de l'American Educational Research Association de San Diego, CA. Documentation disponible à [http://www.education.umd.edu/EDMS/fac/Rupp/].

Rupp, A. A., & Templin, J. (2008). Unique characteristics of diagnostic classification models: A comprehensive review of the current state-of-the-art. *Measurement: Interdisciplinary Research & Perspective, 6*(4), 219-262.

Rupp, A. A., Templin, J., & Henson, R. J. (2010). *Diagnostic measurement: Theory, methods, and applications*. New York : The Guilford Press.

Sinharay, S., & Haberman, S. J. (2009). How much can we reliably know about what examinees know? *Measurement: Interdisciplinary Research & Perspective, 7*(1), 46-49.

Stout, W. (2007). Skills diagnosis using IRT-based continuous latent trait models. *Journal of Educational Measurement, 44*(4), 313-324.

Tatsuoka, C. (2009). Diagnostic models as partially ordered sets. *Measurement: Interdisciplinary Research & Perspective, 7*(1), 49-53.

Tatsuoka, K. K. (1983). Rule-space: an approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20*, 345-354.

Tatsuoka, K. K. (1984). Caution indices based on item response thery. *Psychometrika, 49*(1), 95-110.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. In P. D. Nichols, S. F. Chipman & R. L. Brennan (dir.), *Cognitively diagnostic assessment* (pp. 327-360). Hillsdale, NJ : Erlbaum.

Tatsuoka, K. K. (2009). *Cognitive assessment: An introduction to the rule space method*. New York: Routledge Taylor & Francis Group.

Tatsuoka, K. K., Corter, J. E., & Tatsuoka, C. (2004). Patterns of diagnosed mathematical content and process skills in TIMSS-R across a sample of 20 countries. *American Educational Research Journal, 41*(4), 901-926.

Templin, J., & Henson, R. A. (2005). *The random effects reparametrized unified model: A model for joint estimation of discrete skills and continuous ability*. Princeton, NJ: Educational testing service external research group technical report.

Templin, J., & Henson, R. A. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287-305.

von Davier, M. (2005). *A general diagnostic model applied to language testing data* (Research Report No. RR-05–16). Princeton, NJ : Educational Testing Service.

von Davier, M. (2009). Some notes on the reinvention of latent structure models as diagnostic classification models. *Measurement: Interdisciplinary Research & Perspective, 7*(1), 67-74.

Whitely, S. E. (1980). Multicomponent latent trait models for ability tests. *Psychometrika, 45*, 479-494.

Wilhelm, O., & Robitzsch, A. (2009). Have cognitive diagnostic models delivered their goods? Some substantial and methodological concerns. *Measurement: Interdisciplinary Research & Perspective, 7*(1), 53-57.

Yan, D., Almond, R., & Mislevy, R. (2003). *Empirical comparisons of cognitive diagnostic models*. Princeton, NJ : Educational Testing Service.

Yan, D., Mislevy, R. J., & Almond, R. G. (2003). *Design and analysis in a cognitive assessment* (Research Report No. RR-03–32)). Princeton, NJ: Educational Testing Service.

Yepes-Baraya, M. (1998). *Application of the rule-space methodology to the 1996 NAEP science assessment : grade 4 preliminary results*. Washington, DC: Office of Educational Research and Improvement (ED).

Date de réception : 13 avril 2010

Date de réception de la version finale : 28 janvier 2011

Date d'acceptation : 4 février 2011