Mesure et évaluation en éducation

Quelques nouveaux modèles de mesure

Nathalie Loye

Volume 28, numéro 3, 2005

URI : https://id.erudit.org/iderudit/1087030ar
DOI : https://doi.org/10.7202/1087030ar

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. (2005). Quelques nouveaux modèles de mesure. *Mesure et évaluation en éducation*, *28*(3), 51–68. https://doi.org/10.7202/1087030ar

Résumé de l'article

La combinaison de la psychométrie et de la psychologie cognitive engendre de nouveaux modèles de mesure. Ceux-ci permettent d'établir le portrait de sujets ayant passé un test relativement à une liste d'attributs cognitifs. Cet article présente une comparaison de cinq modèles qui allient la vision psychométrique à la vision cognitive. Les comparaisons portent sur l'étude de la qualité des attributs cognitifs posés a priori, sur les caractérisations des sujets, sur la possibilité d'appliquer les modèles et d'interpréter les résultats et sur les notions de validité et de fidélité.

Quelques nouveaux modèles de mesure

Nathalie Loye
Université de Montréal

Mots clés : Modèles cognitifs, évaluation diagnostique, tests à choix multiples, attributs cognitifs, stratégies de réponse, validité, fidélité

La combinaison de la psychométrie et de la psychologie cognitive engendre de nouveaux modèles de mesure. Ceux-ci permettent d'établir le portrait de sujets ayant passé un test relativement à une liste d'attributs cognitifs. Cet article présente une comparaison de cinq modèles qui allient la vision psychométrique à la vision cognitive. Les comparaisons portent sur l'étude de la qualité des attributs cognitifs posés a priori, sur les caractérisations des sujets, sur la possibilité d'appliquer les modèles et d'interpréter les résultats et sur les notions de validité et de fidélité.

Key words : Cognitive models, diagnostic assessment, multiple-choice tests, cognitive attributes, strategy to answer a question, validity, reliability

New measurement models can be generated through the combination of psychometrics and cognitive psychology. Following a test, these models permit the creation of student portraits in relation to specific cognitive abilities. This paper presents a comparative examination of five models that link both psychometric and cognitive views. Comparisons are made in several areas, including the validity and reliability of the models, as well as how interpretable the results are, how the quality of the attributes is studied a priori, and how students are characterized.

Palavras-chave : Modelos cognitivos, avaliação diagnóstica, testes de escolha múltipla, atributos cognitivos, estratégias de resposta, validade, fidelidade

A combinação da psicometria e da psicologia cognitiva gera novos modelos de medida. A partir de um teste, estes modelos permitem traçar o retrato dos sujeitos, relativamente a uma lista de atributos cognitivos. Este artigo apresenta uma comparação de cinco modelos, que aliam a visão psicométrica à visão cognitiva. As comparações incidem no estudo da qualidade dos atributos cognitivos a priori, nas caracterizações dos sujeitos, na possibilidade de aplicar os modelos e de interpretar os resultados e, ainda, nas noções de validade e de fidelidade.

Note de l'auteure : Toute correspondance peut être adressée par courriel à l'adresse suivante : [nathalie.loye@umontreal.ca].

Introduction

L'évaluation diagnostique est un mode d'évaluation qui a pour but d'apprécier les caractéristiques individuelles d'un sujet (style cognitif, style d'apprentissage, intérêt, motivation, maîtrise des préalables, etc.) et de l'environnement pédagogique, lesquelles devraient avoir des influences positives ou négatives sur son cheminement d'apprentissage (Legendre, 2005, p. 640).

L'évaluation diagnostique en éducation vise à cerner les particularités des élèves en difficulté. Elle cherche à déterminer les forces et les faiblesses des sujets qui passent un test, en s'intéressant à la démarche qui leur permet d'aboutir à la réponse plutôt qu'à la réponse elle-même. Ainsi, l'enjeu d'une telle évaluation consiste à déterminer les processus cognitifs mobilisés afin de prendre des décisions permettant de fournir à chaque sujet en difficulté une remédiation appropriée. Dans des situations complexes et authentiques, l'observation des sujets dans leur résolution de problèmes ou encore la trace de leurs démarches sont autant de façons d'inférer les processus cognitifs mis en œuvre. Toutefois, le temps nécessaire à la mise en place de telles procédures et à leur correction peut justifier le recours à un test diagnostique à correction objective, plus rapide à administrer et à corriger comme peut l'être un examen à choix multiple. Dans la phase d'élaboration d'un tel test, il est important de cibler ce qui doit être évalué, la forme que doit prendre le test, mais également la façon dont les résultats seront interprétés. Les modèles modernes de la mesure qui font l'objet de cet article allient les approches psychométriques et cognitives afin d'inférer, dans un but diagnostique, les processus cognitifs maîtrisés ou non par chaque sujet, même lorsque les seules données accessibles sont la réussite ou l'échec à chaque item. Ces modèles, souvent appelés modèles cognitifs, permettent donc d'exploiter un questionnaire à choix multiple dans la perspective d'établir un diagnostic cognitif individuel.

Psychométrie et sciences cognitives

La psychométrie offre de multiples possibilités pour modéliser les réponses à un test. La théorie classique des tests estime le score vrai des sujets à partir de leurs scores observés. La théorie de la généralisabilité offre de modéliser diverses sources d'erreur. La théorie de réponse à l'item modélise la probabilité de répondre correctement à un item en fonction de variables latentes. Ces variables latentes représentent, par exemple, l'habileté des sujets ou la difficulté des items, estimées à partir des bonnes et des mauvaises réponses

aux items d'un test. L'intérêt d'incorporer la notion d'attribut cognitif dans le cadre d'une évaluation diagnostique est motivé par le souci d'établir les processus cognitifs maîtrisés ou non par chaque sujet, afin de lui fournir ensuite une aide appropriée. Ainsi, ces modèles cherchent à inférer, à partir des réponses à des items, la structure des processus cognitifs, des connaissances et des compétences des étudiants (Nichols, 1994). La différence entre une approche psychométrique, comme celles liées à la théorie classique des tests (TCT) ou encore à la théorie de réponse à l'item (TRI), et une approche cognitive tient dans la finalité. La TCT et la TRI cherchent à estimer l'habileté des individus sur un continuum souvent unidimensionnel de façon à les différencier par une mise en rang. L'approche cognitive cherche plutôt à relever et inférer les différences de structure des individus dans les processus mis en œuvre pour répondre aux questions.

La psychométrie apporte dans le processus les modèles probabilistes, les routines d'estimation des paramètres, une vision quantitative et statistique des processus de réponse. Elle modélise l'aspect continu du comportement du sujet. De son coté, la vision cognitive aborde l'aspect qualitatif et discret des attributs cognitifs définis par les spécialistes de la cognition (DiBello, Stout & Roussos, 1995). Elle offre la possibilité de comprendre la performance liée à la variable latente (Anderson, Greeno, Reder & Simon, 2000). Un attribut cognitif peut être défini de façons diverses par la description de connaissances procédurales ou déclaratives permettant la création d'une stratégie pour répondre à la question, par un état latent de connaissance ou encore par une habileté qu'il n'est pas possible de mesurer directement (Leighton, Gierl & Hunka, 2002 ; Tatsuoka, 1983, 1995 ; Tatsuoka, Birenbaum, Lewis & Sheehan, 1993 ; Tatsuoka & Tatsuoka, 1997). La combinaison de ces deux visions aboutit à une approche mixte, à la fois quantitative et qualitative.

Chaque sujet répond aux items du test, générant ainsi une base de données dans laquelle la valeur 0 symbolise un échec et la valeur 1 une bonne réponse. Aucun des modèles cognitifs actuels ne permet d'utiliser des scores ou des réponses sur une échelle continue, ordinale ou nominale. Ces données peuvent être modélisées grâce à la TRI. L'habileté θ de chaque sujet est alors estimable. Le fait de chercher à établir un portrait cognitif des sujets demande, en plus, de relier chaque item et chaque sujet à un ensemble d'attributs cognitifs. Tous les modèles dont il sera question, dans la suite de cet article, se basent sur l'établissement d'une liste d'attributs cognitifs qu'il est essentiel de maîtriser pour répondre correctement aux items du test (Hartz, 2002 ; Junker, 1999). Le lien entre les attributs cognitifs et les items est opérationnalisé par une matrice

notée Q, basée sur le jugement d'experts et l'analyse des tâches. Cette matrice peut être obtenue dans une approche déductive par les experts lorsque le test est élaboré à partir de la liste des attributs cognitifs. Dans ce cas, un cadre conceptuel contenant la liste des attributs cognitifs permet de générer les items qui sont alors fabriqués en même temps que la matrice. La matrice Q peut également être obtenue après la création du test dans une approche inductive. Dans ce cas, les experts doivent déterminer les attributs cognitifs et leurs liens avec les items, avec la contrainte d'utiliser des items déjà existants. Une telle façon de faire donne la possibilité d'améliorer le pouvoir diagnostique d'un test ou d'exploiter l'information diagnostique d'un examen non initialement élaboré à cette fin. La figure 1 présente un exemple des liens qui peuvent être établis entre trois items et cinq attributs par des experts.

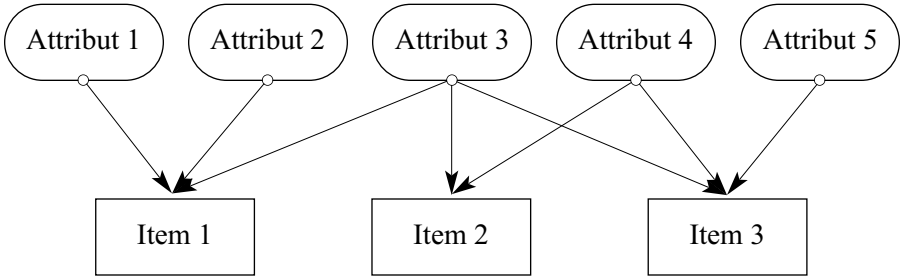

Figure 1. ***Exemple de liens entre des attributs et des items. Les formes arrondies représentent les attributs, qui sont des variables latentes. Les rectangles représentent les réponses des sujets, qui sont des variables observées.***

Dans le cas de cet exemple, la matrice Q, dont les lignes représentent les items et dont les colonnes représentent les attributs, a la forme suivante :

$$Q = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{bmatrix}$$

La fabrication de Q soulève des problèmes quant à la vérification de sa qualité et des différences de diagnostics occasionnées par des matrices différentes : en effet, le jugement des experts intervient sur le plan du choix des attributs mais également sur le plan de leurs liens avec les items du test, et peut varier d'un juge à l'autre. En outre, le fait d'établir une matrice unique suppose de faire un choix lorsqu'un item peut être résolu de diverses façons. Autrement dit, une seule stratégie, comprendre combinaison d'attributs, apparaît dans la matrice pour chaque item. Ainsi, la validité du modèle est remise en cause dès que des sujets utilisent une stratégie différente.

Les modèles cognitifs permettent ensuite de relier les sujets aux attributs. Les paragraphes qui suivent présentent cinq modèles modernes de la mesure combinant les approches psychométriques et cognitives. Ces modèles sont rapidement décrits puis comparés de manière à définir les avantages et les inconvénients à les utiliser dans la perspective d'établir un diagnostic individuel des sujets. Le texte qui suit est organisé en cinq parties :

1. Un aperçu de cinq modèles.
2. Une comparaison relativement à une application de ces modèles à des données, à l'estimation et à l'interprétation des paramètres.
3. Une comparaison de ces modèles relativement à leur façon de caractériser les sujets dans une perspective diagnostique.
4. Une comparaison de ces modèles relativement à la possibilité de juger la matrice Q en rapport avec la validité et la fidélité.
5. Une présentation de leurs avantages et inconvénients quand vient le temps de choisir un modèle.

La partie 5 inclut deux tableaux récapitulatifs regroupant les cinq modèles et est suivie d'une conclusion.

Un aperçu de quelques modèles

Fischer (1973) a été le premier à utiliser une décomposition en attributs de base pour estimer la difficulté des items d'un test. Son *Linear Logistic Test Model* (LLTM) est une extension du modèle unidimensionnel de Rasch (Bechger, Verstralen & Verhelst, 2000 ; DiBello et al., 1995 ; Embretson, 1985a et 1985b, 1999 ; Embretson & Reise, 2000 ; Fisher, 1973 ; Stout, 2002). Ce modèle postule que le paramètre de difficulté d'un item peut être exprimé par une combinaison linéaire d'attributs cognitifs et que tous les sujets utilisent les mêmes processus cognitifs pour résoudre tous les items (Fisher, 1973 ; Sijtsma &

Verweij, 1999). De plus, c'est un modèle compensatoire : un sujet qui possède une habileté élevée sur une composante A et une faible habileté sur une composante B a la même probabilité de répondre correctement à un item qu'un sujet qui possède une faible habileté sur la composante A et une habileté élevée sur la composante B (DiBello et al., 1995).

Embretson (1985a et 1985b) a élaboré une série de modèles complexes qui prennent en compte les attributs cognitifs et qui sont non compensatoires. Le *Multicomponent Latent Trait Model* (MLTM) opérationnalise le concept selon lequel il faut avoir la bonne information sur plusieurs attributs cognitifs pour répondre correctement à un item (Embretson, 1985b ; Embretson & Reise, 2000). Combinant cette approche et celle de Fischer, le modèle *General component latent trait model* (GLTM) tient compte des difficultés des attributs. Ce modèle fournit un outil performant dans une perspective diagnostique car il a pour but l'estimation de la difficulté de l'item mais aussi de l'habileté du sujet relativement à chaque attribut (Embretson, 1993, 1999 ; Embretson & Reise, 2000).

Le modèle *rule space* de Tatsuoka groupe les sujets dans des classes latentes qui reflètent leur état de connaissance (Embretson & Reise, 2000 ; Tatsuoka, 1983). Le sujet ne peut répondre correctement à un item que s'il maîtrise tous les attributs cognitifs qui lui sont reliés. Ce modèle, qui suppose qu'un sujet ayant un schéma particulier d'attributs maîtrisés va les utiliser pour répondre et non pas deviner ou faire des fautes d'étourderie (Tatsuoka, 1983, 1995), est défini pour deux dimensions : (a) le paramètre de compétence du sujet noté θ et (b) un indice représentant la façon de répondre, notée ζ (Embretson & Reise, 2000 ; Tatsuoka, 1983, 1995). À partir de la liste des attributs établie par les spécialistes, l'algèbre de Boole permet de déterminer tous les états de connaissances possibles selon que les attributs sont ou non maîtrisés. Le fait que certains attributs soient des préalables par rapport à d'autres permet de minimiser le nombre d'états de connaissance idéaux, sur la base du jugement des experts et grâce à de nombreux théorèmes (Tatsuoka et al., 1993). Dans un plan cartésien, il est alors possible de représenter graphiquement les couples (θ , ζ) de chaque sujet et de chaque état idéal. C'est la distance de chaque sujet à chaque état idéal qui permet de décider de son classement dans un état particulier. Si le point caractérisant le sujet est trop éloigné de tous les points idéaux, le sujet correspondant peut ne pas être classé (Buck & Tatsuoka, 1998 ; Tatsuoka, 1983).

Le modèle de fusion (*unified model* ou *fusion model*) (DiBello et al., 1995) ajoute la modélisation détaillée de la source des déviations systématiques par rapport au schéma idéal. Un sujet est caractérisé, selon une approche psychométrique, par un trait latent continu θ qui peut être unidimensionnel ou multidimensionnel. De plus, il est supposé, selon une approche cognitive, posséder ou pas chacun des attributs, ce qui revient à le caractériser par un vecteur $\alpha = (\alpha_1, \alpha_2, ..., \alpha_k)$ dans lequel les α_i sont dichotomiques. La probabilité de dévier des schémas établis par la matrice Q est modélisée en fonction de θ, trait latent continu, et de α, vecteur latent discret, là où le *rule space* n'utilise que θ (DiBello et al., 1995 ; Hartz, 2002 ; Stout, 2002). Ce modèle pose le postulat de l'existence d'une structure possédant un certain nombre (inconnu) de classes latentes. L'idée centrale est que dans la réalité, le comportement des sujets sera différent de ce qui a été choisi au départ dans la matrice Q. Quatre sources de variations sont incluses dans le modèle : (a) la stratégie (*strategy*) : choix d'une autre stratégie que celle de Q, (b) le manque d'attributs (*completeness*) : d'autres attributs que ceux de Q peuvent aider à répondre, (c) le fait qu'un attribut soit mal utilisé (*positivity*) : attribut non possédé mais utilisé et attribut possédé mais mal utilisé, et (d) l'erreur aléatoire telle une erreur de transcription ou d'attention. Il est important de comprendre que le fait qu'un attribut soit mal utilisé est plus subtil que le fait de simplement deviner ou de faire une erreur d'étourderie. En effet le partage des sujets en deux catégories, caractérisées par α_i qui peut prendre les valeurs 0 ou 1, amène à considérer que certains sujets peuvent utiliser, par exemple de façon routinière, avec succès un attribut sans nécessairement le maîtriser (α_i=0) ou maîtriser un attribut (α_i=1) mais ne pas répondre correctement, par exemple parce que la question les décourage (Stout, 2002).

La méthode des réseaux de Bayes (*Bayes Net*) a été développée dans la perspective du diagnostic médical (Mislevy, 1995 ; Tatsuoka & Tatsuoka, 1997). Tous les modèles à variables latentes précédents partent du principe que la valeur de l'habileté du sujet implique une distribution de probabilité des scores qui peut se formaliser par $P(X|\theta)$. Toutefois le théorème de Bayes donne l'occasion de calculer $P(\theta|X)$ grâce à la relation $P(\theta|X=x) \sim P(X=x|\theta).P(\theta)$. Il s'agit de définir les distributions de probabilités de toutes les quantités qui interviennent dans le modèle, qu'elles soient observables ou non. Chaque sujet est caractérisé par son état de connaissance sous la forme d'un vecteur latent θ et d'un vecteur de ses scores observés X. Il est possible théoriquement d'estimer $P(\theta|X)$**,** cependant les calculs peuvent très rapidement devenir extrêmement complexes et ne pas aboutir à l'estimation des paramètres si l'on

ne simplifie pas les relations entre les attributs et entre les attributs et les items (Levy & Mislevy, 2004 ; Mislevy, 1993, 1995, 2004 ; Mislevy, Almond, Yan & Steinberg, 2000 ; Mislevy, Steinberg & Russell, 2003). La méthode consiste à relier les attributs et les items selon un réseau tenant compte des relations qui interviennent dans la résolution des items, des relations d'ordre qui existent entre les attributs, des combinaisons d'attributs nécessaires, des diverses stratégies possibles, etc. L'idée générale est que le grand nombre de relations qui existent entre un grand nombre de variables peut être exprimé de façon plus simple par un nombre plus petit de relations entre des variables regroupées (Mislevy et al., 2003).

Application, estimation et interprétation

Medina-Díaz (1993) donne un exemple d'application du modèle LLTM à des données. Les paramètres obtenus sont des indicateurs de la difficulté des attributs cognitifs inclus dans la matrice Q. Ils sont à la fois estimables et interprétables. L'estimation est faite par maximum de vraisemblance.

Les modèles MLTM et GLTM sont des modèles qui font intervenir de nombreux paramètres difficiles à estimer et à interpréter lorsqu'on applique ces modèles à des données (Hartz, 2002). Toutefois, un exemple d'application est proposé par Embretson, Schneider et Roth (1986) sans qu'il soit fait mention de l'utilisation d'un logiciel ou d'un programme spécifique. Le mode d'estimation est également le maximum de vraisemblance.

Le modèle *rule space* présente l'avantage de ne pas avoir de nombreux paramètres à estimer. Il peut facilement être appliqué à des données dès qu'il est possible de définir une liste d'attributs cognitifs liés aux items, même si cette étape est importante et laborieuse (Rupp, 2005). L'interprétation est facilitée par la visualisation graphique de l'espace contenant les points représentant les sujets et les états de connaissance idéaux. L'estimation est faite par maximum de vraisemblance et la distance utilisée pour le classement des sujets est la distance de Mahalanobis.

Le modèle de fusion comporte de nombreux paramètres à estimer puisqu'il modélise l'habileté des sujets sous la forme d'une variable latente continue et leur appartenance à des classes latentes discrètes, en plus de diverses sources de déviation. Les paramètres du modèle original ne sont toutefois pas estimables lorsqu'il est appliqué à des données (Hartz, 2002). Jiang (1996) propose un étalonnage du modèle qui rend les paramètres estimables par maximum de vraisemblance. Toutefois, la méthode proposée implique des

calculs très lourds, limite la taille de la matrice Q utilisable en pratique et perd l'esprit du modèle (Templin, 2004). Hartz (2002) fournit une solution au problème d'estimation des paramètres sous la forme d'un modèle de fusion reparamétrisé dans une approche bayésienne, noté RUM pour *Reparameterized Unified Model*. Ce modèle, qui suppose que chaque sujet répond correctement à chaque item si et seulement s'il maîtrise tous les attributs reliés à cet item, permet à la fois l'estimation des paramètres, même s'ils sont différents de ceux du modèle initial, et leur interprétation. Les paramètres sont alors estimés dans une approche bayésienne par un algorithme Monte Carlo Markov Chain (MCMC).

Enfin l'élaboration du réseau est coûteuse puisqu'elle nécessite la collaboration des experts et des psychométriciens dans le cas du modèle des réseaux de Bayes (Yan, Almond & Mislevy, 2003). Cependant, l'estimation des paramètres ne pose ensuite pas vraiment de problème, pas plus que l'interprétation (Yan et al., 2003). L'approche est bayésienne et l'algorithme MCMC est utilisé. Le modèle de fusion et le modèle des réseaux de Bayes pourraient être combinés pour apporter chacun leurs forces particulières : la force du réseau et la modélisation explicite des sources de variations (Yan et al., 2003).

Les modèles peuvent être classés selon leur facilité d'application à des données. D'un côté, le modèle LLTM fournit des paramètres concernant les items et qui sont estimables en utilisant l'algorithme en FORTRAN de Fischer et Formann (1972). Ces paramètres ne posent pas de problème d'interprétation, ils sont liés à la difficulté des items. De plus, le modèle *rule space* et le logiciel BUGLIB (Tatsuoka, Varadi & Tatsuoka, 1992) estiment des paramètres liés aux sujets et faciles à interpréter. À l'autre extrémité se trouvent les modèles MLTM, GLTM et de fusion, qui sont des modèles dont les paramètres sont difficiles à estimer. En outre, ceux des modèles MLTM et GLTM sont difficiles à interpréter. Le modèle RUM et le modèle des réseaux de Bayes offrent des résultats qui peuvent être comparables en ce qui concerne le diagnostic des sujets (Yan et al., 2003). Enfin, le modèle de fusion reparamétrisé (RUM) est le seul modèle à proposer des paramètres estimables et interprétables pour juger de la qualité des attributs. Le logiciel Arpeggio, propriété de l'*Educational Testing Service*, permet l'application du modèle RUM à des données et l'estimation des paramètres (Stout, 2002). L'application du modèle des réseaux de Bayes peut se faire grâce au logiciel BUGS (Spiegelhalter, Thomas, Best & Gilks, 1995).

Une comparaison en ce qui concerne la caractérisation des sujets

Le modèle LLTM caractérise les sujets de la même façon que le modèle de Rasch de la TRI, dont il est issu. Il permet d'estimer la valeur de leur habileté sous forme d'une variable latente continue, notée θ.

Les modèles MLTM et GLTM caractérisent les sujets sous la forme d'un vecteur d'habiletés, variables latentes continues, pour chacun des attributs de la liste de départ. Ces modèles offrent un diagnostic pour chaque sujet en fonction de chacun des attributs.

Le modèle *rule space* caractérise chaque sujet par un couple (θ , ζ). L'estimation de ce couple de coordonnées n'est qu'une étape visant la classification de chaque sujet dans un état de connaissance idéal. Ce dernier est représenté par un vecteur discret de la forme $\alpha = (\alpha_1, \alpha_2, ..., \alpha_k)$ dans lequel k est le nombre d'attributs et α_i peut prendre les valeurs 0 ou 1. La particularité vient du fait que le vecteur est obtenu par classement plutôt que par estimation. Le diagnostic individuel se fait de façon qualitative en énonçant quels attributs sont maîtrisés ou non par chaque sujet.

Le modèle de fusion et le RUM utilisent à la fois θ et $\alpha = (\alpha_1, \alpha_2, ..., \alpha_k)$ pour calculer les probabilités de dévier des réponses prédites par Q. Ainsi, chaque sujet est caractérisé par la valeur de son habileté estimée sur une échelle continue, comme dans les modèles de la TRI, et par un vecteur $\alpha = (\alpha_1, \alpha_2, ..., \alpha_k)$ qui est alors estimé. Le diagnostic peut se faire de la même façon que précédemment par la liste des attributs maîtrisés et non maîtrisés par chaque sujet.

Enfin, le modèle des réseaux de Bayes permet de caractériser chaque sujet par l'estimation d'un vecteur $\alpha = (\alpha_1, \alpha_2, ..., \alpha_k)$ pour un diagnostic sous forme de la liste des attributs maîtrisés ou non.

Le diagnostic est donc possible sous des formes comparables dans les modèles *rule space*, de fusion, de fusion reparamétrisé et des réseaux de Bayes, même si les procédures varient d'un modèle à l'autre.

Validité et fidélité de la matrice Q

La plupart des modèles cognitifs permettent d'inférer les processus cognitifs mis en œuvre par les sujets pour répondre aux items en se basant sur la matrice Q qui lie items et attributs. De la validité et de la fidélité de la matrice Q dépendent la validité et la fidélité des mesures produites et donc du diagnostic individuel. Dans le cas du modèle LLTM (Fischer, 1973), la matrice Q ne joue un rôle que dans la description de la difficulté des items car le

modèle ne permet pas de relier les attributs aux sujets. Aucune procédure incluse dans le modèle n'offre de tester la qualité de la matrice Q. Medina-Díaz (1993) utilise une analyse factorielle confirmatoire pour valider la structure cognitive de la matrice Q. Dans le même esprit, Dimitrov et Raykov (2003) proposent deux méthodes externes de validation de la matrice Q pour ce modèle. La première consiste à comparer la difficulté des items obtenue grâce au modèle de Rasch à la difficulté des items obtenue par le modèle LLTM. La deuxième méthode consiste à établir un cheminement entre les items (*path diagram*) basé sur la matrice Q et sur l'hypothèse qu'un item faisant appel à certains attributs sera relié de façon unidirectionnelle à un autre item faisant appel aux même attributs associés à un ou plusieurs autres (Dimitrov, 1994). De cette façon, tous les items peuvent former un réseau hiérarchique qui peut être testé par une analyse structurelle pour données dichotomiques utilisant la matrice de corrélation tétrachorique des réponses aux items. L'ajustement des données au modèle permet alors de valider la matrice Q. Enfin, les items sont supposés avoir une difficulté croissante au fur et à mesure de la progression le long du réseau hiérarchique (Dimitrov, 1994). La vérification de cette hypothèse permet également de valider les attributs et leurs liens avec les items.

Les modèles MLTM et GLTM sont des modèles complexes qui permettent de caractériser les sujets et les items, mais avec l'hypothèse que chaque attribut est requis par chaque item. En d'autres termes, cela signifie que la matrice Q ne comporte que des 1. Ils ne permettent donc pas de juger de la pertinence des liens entre les items et les attributs dans le même sens que les autres modèles.

Le modèle *rule space* groupe les sujets dans des classes latentes discrètes. Ce modèle n'offre pas directement la possibilité de s'assurer de la qualité de la matrice Q car toute variation par rapport au modèle est classée comme erreur aléatoire. Toutefois, le taux de classification des sujets dans les états idéaux est un indicateur de la qualité de la matrice (Buck & Tatsuoka, 1998; Buck, Tatsuoka & Kostin, 1997). Ainsi, en procédant à plusieurs analyses successives, il est possible de raffiner la liste des attributs et d'aboutir à une matrice qui permet de classer un pourcentage de sujets jugé satisfaisant, par exemple 80% et plus selon Buck et ses collègues. De ce point de vue, même si le modèle ne donne pas la possibilité de juger de la qualité des choix des attributs, le fait de l'appliquer plusieurs fois de suite, avec des matrices améliorées à chaque étape, permet d'avoir une certaine confiance dans le choix de la matrice Q. De plus, diverses analyses statistiques, telles des régressions multiples cherchant à prédire la difficulté des items ou des analyses de corrélation entre les attributs donnent la

possibilité d'éliminer ceux qui sont peu corrélés avec la difficulté des items et ceux qui sont trop corrélés avec les autres attributs (Buck & Tatsuoka, 1998). Ces diverses analyses ont pour but une amélioration et une validation de la qualité de la liste des attributs finalement retenus dans un processus itératif.

Le modèle de fusion permet théoriquement d'estimer des paramètres concernant le choix de la stratégie de chaque sujet, le manque d'attributs (*completeness*), la mauvaise utilisation des attributs (*positivity*) et l'erreur aléatoire. Il fournit donc au chercheur le moyen de juger de la qualité de Q. Le fait qu'un sujet puisse utiliser une autre stratégie que celle proposée par Q ou que des attributs ne soient pas présents dans la matrice peut être, au moins en partie, attribué à un manque de validité du modèle (Junker, 1999). En effet, si les attributs proposés ne correspondent pas à la réalité du comportement cognitif des sujets, la validité du modèle est remise en cause (Corter, 1995). D'un autre côté, le fait qu'un attribut soit mal utilisé (*positivity*) est associé à la notion de fidélité (Junker, 1999). Le modèle RUM (Hartz, 2002) est une version modifiée du modèle de fusion pour lequel les paramètres, qui sont toutefois partiellement différents, sont estimables. Les nouveaux paramètres sont au nombre de deux et sont notés π_i* et π_{ik}*. Le paramètre π_i* représente la probabilité qu'un sujet ayant maîtrisé tous les attributs requis pour l'item i par la matrice Q les ait convenablement utilisés pour répondre à la question. Ce paramètre peut être associé à la notion de fidélité des attributs. Le paramètre π_{ik}* compare la probabilité de bien répondre à l'item i selon que le sujet maîtrise ou non l'attribut k. Ce paramètre représente la pénalité due au fait de ne pas maîtriser l'attribut ; il peut être associé à la notion de validité puisqu'il est un indice de la pertinence des attributs. Enfin, le paramètre présent dans le modèle original jugeant si la matrice Q contient tous les attributs importants est conservé et est un indicateur de la validité du modèle. Ainsi le modèle RUM garde l'esprit du modèle original tout en permettant son application à des données et l'estimation de ces paramètres qui peut être faite dans une approche bayésienne et non fréquentiste. Le modèle de fusion dans sa forme RUM est donc le seul des modèles présentés qui offre des paramètres estimables et interprétables pour ce qui est de la validité et de la fidélité de la matrice Q.

Dans le cas des réseaux de Bayes, le réseau est réalisé par les experts à fois théoriquement et empiriquement (Mislevy et al., 2003) et relie les items aux attributs mais aussi les items et les attributs entre eux. Le travail que demande sa fabrication prend appui sur l'expérience de ses créateurs mais aussi sur les distributions de probabilités relatives à chaque nœud du réseau. Cette façon de faire fournit une validation de la matrice Q.

Ce qui précède permet donc de conclure que le problème de la qualité de la matrice Q est soulevé et abordé différemment par les divers modèles. Les méthodes de validation sont externes au modèle dans le cas du *rule space* ou du LLTM. Elles sont explicitement inclues dans le modèle de fusion et dans sa version reparamétrisée alors que la création du réseau de Bayes s'appuie sur des analyses statistiques permettant de la valider. Les modèles MLTM et GLTM ne permettent pas de valider les liens entre les items et les attributs puisque tous les attributs sont reliés à tous les items.

Le choix d'un modèle

Selon Di Bello et ses collègues (1995) et Junker (1999), un modèle plus fin permet d'assurer une meilleure fidélité ou validité au prix d'une faible maniabilité, alors qu'un modèle plus grossier est plus facile à appliquer à des données. Tous les modèles utilisent des données dichotomiques. Parmi les cinq modèles présentés, le LLTM ne relie pas les sujets aux attributs. Les quatre autres modèles tiennent plus ou moins compte des nombreux facteurs qui interviennent dans le processus de réponse des sujets. Les modèles MLTM, GLTM, de fusion et de fusion reparamétrisé sont les modèles qui assurent la meilleure validité car ils collent le plus à la réalité du processus de réponse. Le modèle *rule space* est le modèle le plus grossier mais le plus facile à appliquer. Le tableau 1 présente une synthèse des comparaisons faites entre ces modèles alors que le tableau 2 présente les avantages et inconvénients de chacun d'eux.

Tableau 1
Récapitulatif des comparaisons entre les modèles

Modèle	*Référence initiale*	*Existence de paramètres pour juger la qualité de Q*	*Caractérisation des sujets*	*Estimation et interprétation des paramètres*	*Méthode d'estimation des paramètres*	*Logiciel*
LLTM	Fischer, 1973	Non	θ	Oui	Maximum de vraisemblance	Algorithme en FORTRAN de Fischer et Formann (1972)
MLTM GLTM	Embretson, 1985a, 1985b	Non	θ (sur chaque attribut)	Oui, sous certaines conditions, interprétation complexe	Maximum de vraisemblance	-
Rule space	Tatsuoka, 1983	Non	(θ , ζ) et (α_i, ...α_k) (classement)	Oui, interprétation aisée	Maximum de vraisemblance	BUGLIB
Fusion	DiBello, Stout & Roussos, 1995	Oui	θ et (α_i, ...α_k) (estimation)	Non	Maximum de vraisemblance	-
RUM	Hartz, 2002	Oui	θ et (α_i, ...α_k) (estimation)	Oui, interprétation aisée	Monte Carlo Markov Chain MCMC	Arpeggio MCMC
Réseaux de Bayes *(Bayes net)*	Mislevy, 1995	Non	θ et (α_i, ...α_k) (estimation)	Oui, interprétation aisée	Monte Carlo Markov Chain MCMC	BUGS

Note. Les paramètres α_i sont dichotomiques et k est le nombre total d'attributs.

Tableau 2
Avantages et inconvénients des modèles

Modèle	*Avantages*	*Inconvénients*
LLTM	• Caractérise les items • Paramètres estimables et interprétables • Existence d'un programme informatique	• Ne caractérise pas les sujets en fonction des attributs • Ne permet pas de juger de la qualité des attributs
MLTM GLTM	• Modèles à grain fin • MLTM caractérise les sujets • GLTM caractérise les sujets et les items	• Modèles complexes • Paramètres difficiles à estimer et à interpréter • Pas de logiciel existant
Rule space	• Modèle facile à appliquer à des données • Paramètres faciles à interpréter • Existence d'un logiciel	• Ne caractérise pas les items en fonction des attributs • Ne permet pas de juger de la qualité des attributs
Fusion	• Modèle à grain fin • Caractérise les sujets et les items • Paramètres faciles à interpréter • Permet de juger la qualité de Q quant à la validité et à la fidélité	• Modèle complexe • Paramètres non estimables
RUM	• Modèle à grain fin • Caractérise les sujets et les items • Existence d'un logiciel • Paramètres estimables • Paramètres faciles à interpréter • Permet de juger la qualité de Q quant à la validité et à la fidélité	• Modèle complexe
Réseaux de Bayes *(Bayes net)*	• Caractérise les sujets • Existence d'un logiciel • Paramètres estimables • Paramètres faciles à interpréter • La validation des attributs se fait lors de la construction du réseau	• Modèle complexe • Réseau complexe à établir • Ne caractérise pas les items • Pas de paramètres pour juger de la qualité des attributs

Le choix d'un modèle est fonction des besoins des chercheurs et du contexte d'utilisation. Si le but est de définir la structure cognitive d'un test, le choix se portera sur le modèle LLTM qui estime la contribution des attributs à la difficulté de chaque item. Les quatre autres modèles visent le diagnostic individuel des sujets en fonction des attributs cognitifs contenus dans la matrice Q.

Ainsi le choix de l'un ou l'autre des modèles doit être fonction (a) du fait de vouloir caractériser les items ou les sujets, (b) du fait de chercher à étudier ou non la validité et la fidélité de la matrice Q, (c) du fait de préférer un modèle plus simple et moins fin ou un modèle plus complexe et plus fin selon les questions de recherche. Enfin, la facilité à se procurer l'un ou l'autre des logiciels peut également jouer un rôle dans le choix.

Conclusion

Les quelques nouveaux modèles de mesure présentés dans cet article combinent les approches psychométriques et cognitives. Dans le contexte de l'évaluation diagnostique, le gain à utiliser cette double approche est un diagnostic sous la forme de commentaires personnels à chaque élève plutôt que sous celle d'une note ou même d'une estimation numérique de son habileté. Cela offre l'avantage d'envisager une aide sur mesure pour les sujets concernés. De plus, ces modèles permettent le développement de *testing* adaptatif. Cependant, les chercheurs sont mis au défi d'améliorer la création de la matrice Q, de la valider, d'étudier les conditions visant à améliorer l'estimation des paramètres des modèles plus complexes ou encore de définir les liens entre les valeurs des paramètres estimés et les interprétations.

RÉFÉRENCES

Anderson, J.R., Greeno, J.G., Reder, L.M., & Simon, H.A. (2000). Perspectives on learning, thinking, and activity. *Educational Researcher*, *29*, 11-13.

Bayesian inference using Gibbs sampling, Version 0.50. Cambridge : MRC.

Bechger, T.M., Verstralen, H.H.F.M., & Verhelst, N.D. (2000). Equivalent logistic test models. RetD report 2000-4, Arnhem : Cito.

Biostatistics Unit.

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., & 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.

Corter, J.E. (1995). Using clustering methods to explore the structure of diagnostic tests. In P.D. Nichols, S.F. Chipman & R.L. Brennan (éd.), *Cognitively diagnostic assessment* (pp. 306-326). Hillsdale, NJ : Erlbaum.

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 (éd.), *Cognitively diagnostic assessment* (pp. 361-389). Hillsdale, NJ : Erlbaum.

Dimitrov, D.M. (1994). *Hierarchical latent trait approach in test analysis*. Paper presented at the Annual Meeting of the Midwestern Educational Research Association, Chicago, October, 1994.

Dimitrov, D.M., & Raykov, T. (2003). Validation of cognitive structures : a structural equation modeling approach. *Multivariate Behavioral Research, 38*(1), 1-23.

Embretson, S.E. (1985a). Introduction to the problem of test design. In S.E. Embretson (éds), *Test design* : *Developments in psychology and psychometrics* (pp. 3-17). New York : Academic Press.

Embretson, S.E. (1985b). Multicomponent latent trait models for test design. In S.E. Embretson (éds), *Test design* : *Developments in psychology and psychometrics* (pp. 279-294). New York : Academic Press.

Embretson, S.E. (1993). Psychometric models for learning and cognitive processes. In N. Fredericksen, R.J. Mislevy & I. Bejar (éds), *Test theory for a new generation of tests* (pp. 125-150). Hillsdale, NJ : Erlbaum.

Embretson, S.E. (1999). Cognitive psychology applied to testing. In F.T. Durso, R.S. Nickerson, R.W. Schvaneveldt, S.T. Dumais, D.S. Lindsay & M.T.H. Chi (éds), *Handbook of applied cognition* (pp. 629-658). New York : Willey.

Embretson, S.E., Schneider, L.M., & Roth, D.L. (1986). Multiple processing strategies and the construct validity of verbal reasoning tests. *Journal of Educational Measurement, 23*, 13-32.

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

Fischer, G.H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*, 359-374.

Fischer, G.H., & Formann, A.K. (1972). *An algorithm and a FORTRAN program for estimating the item parameters of the linear logistic test model*. Research Bulletin, 24, University of Vienna, Institute of Psychology.

Hartz, S. (2002). *A Bayesian framework for the unified model for assessing cognitive abilities : Blending theory with practicality*. Thèse de doctorat non publiée, University of Illinois, Urbana-Champaign.

Jiang, H. (1996). *Applications of computational statistics in cognitive diagnosis and IRT modeling*. Thèse de doctorat non publiée, University of Illinois, Urbana-Champaign.

Junker, B. (1999). *Some statistical models and computational methods that may be useful for cognitively-relevant assessment*. Rapport de recherche préparé pour le Committee on the Foundations of Assessment.

Legendre, R. (2005). *Dictionnaire actuel de l'éducation* (3e édition). Montréal : Guérin.

Leighton, J.P., Gierl, M.J., & Hunka, S.M. (2002). *The attribute hierarchy model for cognitive assessment*. Paper presented at the annual meeting of NCME, 2-4 avril 2002.

Levy, R., & Mislevy, R.J. (2004). *Specifying and refining a measurement model for a simulation-based assessment.* National center for research on evaluation : rapport de recherche CSE Report 619.

Medina-Díaz, M. (1993). Analysis of cognitive structure using the Linear Logistic Test Model and quadratic assignment. *Applied Psychological Measurement, 17*(2), 117-130.

Mislevy, R.J. (1993). Foundations of a new test theory. In N. Frederiksen, R.J. Mislevy & I. Bejar (éds), *Test theory for a new generation of tests.* Hilldale, NJ : Lawrence Erlbaum Associates, Publishers.

Mislevy, R.J. (1995). Probability-based inference in cognitive diagnosis. In P.D. Nichols, S.F. Chipman & R.L. Brennan (éd.), *Cognitively diagnostic assessment* (pp. 43-71). Hillsdale, NJ : Erlbaum.

Mislevy, R.J. (2004). *A brief introduction to evidence-centered design*. National center for research on evaluation : rapport de recherche CSE Report 632.

Mislevy, R.J., Almond, R.G., Yan, D., & Steinberg, L.S. (2000). *Bayes Nets in educational assessment : where do the numbers come from ?* National center for research on evaluation : rapport de recherche CSE Report 518.

Mislevy, R.J., Steinberg, L.S., & Russell, G.A. (2003). On the structure of educational assessments. *Measurement : Interdisciplinary Research and Perspectives*, *1*(1), 3-62.

Nichols, P.D. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64*(4), 575-603.

Rupp, A.A. (2005). *The answer is in the question : A guide for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models*. Manuscript submitted for publication.

Sijtsma, K., & Verweij, A.C. (1999). Knowledge of solution strategies and IRT modeling of items for transitive reasoning. *Applied Psychological Measurement*, *23*(1), 55-68.

Spiegelhalter, D.J., Thomas, A., Best, N.G., & Gilks, W.R. (1995). *BUGS : Bayesian inference using Gibbs sampling, Version 0.50*. MRC Biostatistics Unit, Cambridge.

Stout, W. (2002). Psychometrics : From practice to theory and back. *Psychometrica*, *67*(4), 485-518.

Tatsuoka, C.M., Varadi, F., & Tatsuoka, K.K. (1992). *BUGLIB*. Unpublished computer program, Trenton, NJ.

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. (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 (éd.), *Cognitively diagnostic assessment* (pp. 305-326). Hillsdale, NJ : Erlbaum.

Tatsuoka, K.K., Birenbaum, M., Lewis, C., & Sheehan, K.M. (1993). Proficiency scaling based on conditional probability functions for attributes. *Educational Testing Service*. Princeton, NJ.

Tatsuoka, K.K., & Tatsuoka, M.M. (1997). Computerizing cognitive diagnostic adaptative testing : Effect on remedial instruction as empirical validation. *Journal of Educational Measurement, 34*(1), 3-20.

Templin, J.L. (2004). *Generalized linear mixed proficiency models for cognitive diagnosis.* Thèse de doctorat non publiée, University of Illinois, Urbana-Champaign.

Yan, D., Almond, R., & Mislevy, R. (2003). *Empirical comparisons of cognitive diagnostic models*. ETS, NJ : Princeton. Article non publié.