Milanovitch, M. (1998). *Studies in Language testing : Multilingual glossary of language testing terms*, Cambridge University Press.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14(2), 97-116.

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling*, 3(1), 3-24.

Wright, B. D. et Bell, S. R. (1984). Items banks : What, why, how. *Journal of Educational Measurement*, 21(4), 331-354.

Wright, B. D. et Linacre, J. M. (1997). *Bigsteps 2.71*. Chicago : MESA Press. Logiciel gratuit et téléchargeable sur le site http ://www.winsteps.com/bigsteps.htm

Wright, B. D., Linacre, J. M., Gustafson, J-E. et Martin-Löf, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, 8(3), 370.

Wright, B. D. et Stone, M. H. (1979). *Best Test Design*. Chicago : MESA Press.

Psychologie et Psychométrie. Vol. 24 - N° 2/3 - 2003

Le testing adaptatif par ordinateur dans la mesure en éducation : potentialités et limites

Romain Martin

Summary : *Computerized adaptive testing (CAT) is a diagnostic tool which raises a growing interest following a larger availability of powerful computer equipment. This article gives a brief historical overview on CAT-development and describes its components, logic and concrete implementation. It will then focus on the potential of CAT-use as a tool for summative and formative evaluation in an educational context. For summative evaluation, large-scale CAT programs in a licensure and certification context as well as specific problems associated with this approach will be presented. For formative evaluation, the question will be asked whether CAT can be used to guide pedagogical and remediative interventions in a learning context. More specifically CAT-use in mastery and in diagnostic testing will be considered. To conclude, the potential benefit of new, more multimedia-based item formats and of connecting CAT computer-platforms in networks will be discussed.*

Key words : *Computerized adaptive testing (CAT) ; educational measurement ; diagnostic testing ; summative and formative evaluation.*

Résumé : *Le testing adaptatif par ordinateur (TAO) est un dispositif de mesure qui connaît un intérêt croissant suite à une disponibilité de plus en plus importante d'un matériel informatique performant. Cet article va donner un bref*

* Université du Luxembourg, B.P.2, L-7201 Walferdange (G.D. de Luxembourg), GRAPCO, Université Nancy 2.

aperçu historique sur le développement du TAO pour ensuite présenter une description succincte des composantes, de la logique de fonctionnement et de la mise en œuvre concrète du TAO. Il va alors se centrer sur les possibilités de mise en œuvre du TAO en tant qu'outil d'évaluation sommative et formative dans un contexte scolaire. Pour l'évaluation sommative, on présentera la mise en œuvre du TAO comme outil de certification à grande échelle en évoquant les problèmes qui en découlent. Pour l'évaluation formative, on analysera dans quelle mesure le TAO peut se prêter à guider des interventions pédagogiques ou des interventions remédiatives auprès d'apprenants. Dans ce contexte, on présentera les possibilités d'utilisation du TAO dans l'apprentissage de maîtrise et comme outil d'évaluation fournissant un diagnostic du fonctionnement cognitif de l'apprenant. Pour conclure, il sera discuté dans quelle mesure des nouveaux formats d'item utilisant plus amplement les fonctionnalités multimédia de l'ordinateur sont susceptibles d'être mis à profit dans une logique de TAO. On analysera également la nécessité d'une mise en réseau des ordinateurs pour une utilisation efficace du TAO dans un contexte éducatif.

Mots clés : *Testing adaptatif par ordinateur (TAO) ; mesure éducationnelle ; testing de maîtrise ; testing diagnostique ; évaluation sommative et formative.*

INTRODUCTION

La société de l'information dans laquelle nous vivons en ce début du 21e siècle se caractérise avant tout par une omniprésence d'ordinateurs devenus très performants et permettant l'utilisation de ressources de traitement de l'information puissantes pour la réalisation de tâches quotidiennes. Dès l'introduction des premiers micro-ordinateurs au courant des années 70, il était visible qu'il allait y avoir une évolution rapide de cette technologie vers des machines de plus en plus performantes accessibles à un public toujours plus large. Aussi est-il compréhensible qu'on prévoyait à l'époque des véritables révolutions dans les domaines qui étaient susceptibles de profiter d'une mise à disposition d'une puissance de calcul importante. Une telle révolution était également annoncée pour les domaines de l'éducation en général, et plus spécifiquement pour celui de la mesure en éducation (Bunderson, Inouye & Olsen, 1989 ; Holtzman, 1970). En ce qui concerne les applications pédagogiques, on voyait dans l'ordinateur un excellent outil pour individualiser les apprentissages scolaires par le développement de logiciels permettant aux élèves de s'auto-former. Or cette formation individualisée était censée être pilotée par un enseignant dont le rôle passait de celui d'un transmetteur de savoirs à celui d'un gestionnaire du processus d'apprentissage, d'où l'importance de disposer d'outils d'évaluation efficaces et faciles à mettre en œuvre. Là encore, l'ordinateur semblait être l'outil de choix, puisque sa puissance de calcul permettait d'envisager des tests qui seraient courts, mais néanmoins fiables du fait d'un choix des items adapté au niveau de compétence des sujets : l'idée du testing adaptatif par ordinateur (TAO) était née (Lord, 1970 ; 1971a; 1971b).

Or, plus de 30 ans après les premiers travaux de Lord, il faut bien constater qu'on dispose de la puissance de calcul qui était souhaitée à l'époque, mais que malgré des développements majeurs en la matière, la révolution annoncée (et en partie déjà constatée) par des auteurs comme Bunderson et al. (1989) n'a pas véritablement eu lieu, ni dans le domaine des apprentissages, ni dans celui du testing assisté par ordinateur. Le présent article se propose donc de faire le bilan des premières applications concrètes du TAO à des problématiques en rapport avec l'éducation qui ont émergé au courant des années 90, en soulignant les problèmes de mise en pratique qui sont apparus et qui expliquent le succès relatif du TAO. Nous allons ensuite faire état des développements récents en matière de TAO, surtout en ce qui concerne son utilisation en tant qu'outil d'évaluation formative dans un contexte scolaire pour enfin donner quelques pistes de développement futur.

LE PRINCIPE DE FONCTIONNEMENT DU TAO

Même si le présent article ne se propose pas de réaliser une présentation exhaustive et détaillée de tous les aspects d'un système de TAO (cf. à ce sujet par exemple Dechef & Laveault, 1999 ; Hulin, Drasgow & Parsons, 1983 ; Wainer, Dorans, Flaugher, Green, Mislevy, Steinberg & Thissen, 2000), il paraît néanmoins indispensable de donner un aperçu succinct des ses différentes composantes. Comme pour tout test adaptatif, une caractéristique essentielle du TAO consiste dans le fait que les degrés de difficulté des items présentés à un sujet particulier sont choisis de manière à être adaptés au niveau de compétence de ce sujet. Un autre élément adaptatif du TAO est le fait qu'on va adapter la longueur du test de manière a faire passer seulement le nombre d'items nécessaires pour atteindre un objectif prédéterminé de qualité de mesure (qui s'exprime en général à travers

l'erreur de mesure). Le but est donc d'arriver à une individualisation de la passation d'un test qui permette d'utiliser exclusivement les items les mieux adaptés pour déterminer efficacement et précisément le niveau de compétence du sujet. Or ce caractère adaptatif a pour conséquence que des sujets différents ne vont pas passer les mêmes items, ce qui pose le problème de la comparabilité des résultats obtenus. En effet, le score brut qui est utilisé dans la théorie classique du score vrai n'est plus utilisable dans le cadre du testing adaptatif, puisque le choix adaptatif des items va normalement impliquer que tous les sujets vont réussir environ la moitié des items. La mise en œuvre efficace du testing adaptatif nécessite donc un modèle de mesure qui permette de situer les sujets sur une même dimension de compétence à partir de résultats obtenus sur des échantillons d'items différents. Un tel modèle de mesure peut être trouvé dans les modèles de réponse à l'item (MRI, cf. Hulin *et al.*, 1983 ; Lord & Novick, 1968). L'étalonnage d'un ensemble d'items à l'aide des MRI implique la détermination, pour chaque item, d'une fonction de réponse à l'item exprimant la relation entre la position du sujet sur un trait latent (en général désigné par la lettre grecque θ) et sa probabilité de réussite à l'item. Cette fonction de réponse est caractérisée par trois paramètres : la difficulté qui précise la position de la fonction sur le trait latent θ, la discrimination et une caractéristique qui rend compte de la probabilité de deviner la bonne réponse dans des questions à choix multiple. Pour l'application des MRI comme modèle de mesure, il faut disposer d'un ensemble d'items qui soient homogènes au sens des MRI, c'est-à-dire pour lesquels une détermination des paramètres des fonctions de réponse ait été possible en respectant toutes les contraintes du modèle (cf. Dickes, 1999). Pour le TAO, on parle en général de banque d'items pour désigner la partie du système qui réalise le stockage des items, comportant à la fois leur contenu et leurs paramètres (en général les paramètres des MRI). Il est alors possible d'estimer à partir du patron de réponses données par le sujet sur un échantillon de ces items sa position sur la dimension latente θ. Les algorithmes les plus couramment employés pour l'estimation de la compétence sont l'algorithme du maximum de vraisemblance (cf. Hulin *et al.*, 1983), ainsi que la méthode bayesienne de Owen (1975). Ces algorithmes impliquent la détermination d'une courbe de vraisemblance qui exprime la probabilité du patron de réponses spécifique qui a été observé chez un sujet particulier en fonction de la dimension de compétence θ. On va alors attribuer au sujet le niveau θ pour lequel cette fonction de vraisemblance atteint son maximum, c'est-à-dire que ce niveau de compétence présente la plus grande probabilité de générer le patron de réponses qui a été observé. Il reste alors encore à clarifier la gestion du caractère adaptatif dans le TAO, c'est-à-dire le choix des items et le critère de fin d'examen. Là encore, on fait le plus souvent appel à des éléments qui sont directement reliés aux MRI. Pour le choix des items, on adopte le plus souvent la règle du maximum d'information qui consiste à choisir comme item suivant celui qui fournit le maximum d'information au niveau de compétence estimée actuelle du sujet (le concept d'information étant défini dans le cadre des MRI). Pour déterminer la fin de la séance d'examen, on se donne en général un critère d'arrêt relatif au degré de précision visé par la mesure. Ce critère peut facilement être quantifié dans le cadre des MRI qui permettent de déterminer l'erreur de mesure associée à une estimation de compétence donnée. On va donc arrêter le test si la précision voulue est atteinte ou si la banque d'items ne fournit plus d'items susceptibles de faire diminuer l'erreur de mesure.

Psychologie et Psychométrie. Vol. 24 - N° 2/3 - 2003

La mise en œuvre du TAO dans un contexte éducatif

Rappelons que le développement de tests pour la mesure du fonctionnement cognitif était dès son origine étroitement lié à des préoccupations éducatives. En effet, le test de Binet-Simon de 1908 qui est considéré comme le premier test d'intelligence était développé comme outil de mesure dans le cadre de l'orientation des élèves vers des enseignements spécialisés. La grande importance de l'évaluation dans le contexte scolaire renvoit d'ailleurs à une double fonction de l'école qui est à la fois une institution éducative et formative, et aussi une instance de sélection, d'orientation et de certification (Ingenkamp, 1989). Cette double fonction se retrouve également au niveau de la mesure et de l'évaluation dans le domaine de l'éducation pour laquelle on fait très souvent la distinction entre une évaluation formative plus centrée sur la fonction éducative de l'école et une évaluation sommative plus centrée sur sa fonction certificative et sélective. Ainsi l'évaluation formative prend une optique essentiellement prospective et processuelle en essayant de déterminer la position d'un apprenant dans un cursus d'apprentissage, l'objectif étant un guidage efficace du processus d'apprentissage par des interventions pédagogiques adaptées au niveau de connaissances actuelles de l'élève (Rouiller, 1998). L'évaluation sommative adopte par contre une optique plus rétrospective et statique avec l'objectif de constater des apprentissages qui ont pu être acquis dans le passé et de sanctionner ces apprentissages par une forme de certification ou de note, sans pour autant qu'il y ait un but explicite de remédiation ou de développement. Or on peut remarquer que si on parle de la mesure en éducation, c'est très souvent la fonction sommative et certificative qui est accentuée au détriment de la fonction formative. Ceci est certainement dû au fait que c'est surtout dans des évaluations certificatives à enjeu important que le souci de fidélité et d'équité de la mesure réalisée est primordial. Cela implique alors le recours à des mesures standardisées nécessitant la mise en œuvre d'une logistique importante qu'il n'est souvent pas possible de déployer dans le contexte des apprentissages scolaires quotidiens. C'est d'ailleurs cette dernière contrainte de mise en œuvre d'une évaluation formative dans le contexte scolaire normal sans la présence de spécialistes de la mesure qui constitue une des plus grandes difficultés dans le développement d'outils d'évaluation formative qui soient basés sur les derniers développements théoriques dans le domaine de la

psychométrie, mais qui puissent également être opérationnels dans un contexte professionnel.

En ce qui concerne la mise en œuvre du TAO dans un contexte éducatif, on a également pu constater cette accentuation de l'évaluation sommative et certificative au détriment d'une évaluation formative dont la mise en œuvre paraît nettement plus difficile. Parmi les quatre générations de mesure éducationnelle assistée par ordinateur qui sont présentées par Bunderson et al. (1989) on constate ainsi que ce sont seulement les dispositifs de TAO les plus évolués qui fournissent un plus grand nombre d'informations utilisables dans une optique d'évaluation formative. Selon ces auteurs, la forme la plus simple de testing assisté par ordinateur consiste dans la transposition de tests papier-crayon sur un support informatisé, une telle transposition ayant été réalisée à maintes reprises jusqu'à aujourd'hui. Le TAO tel qu'il a été décrit plus haut constitue une deuxième génération de tests informatisés dont le caractère adaptatif constitue l'atout majeur, puisqu'il va permettre de raccourcir considérablement les tests sans perdre en précision de mesure. Néanmoins le résultat de la mesure reste toujours un positionnement statique de l'individu sur une dimension latente qui se prête bien à juger si un niveau prédéterminé de performance a pu être atteint par le sujet, mais qui ne permet pas forcément de diagnostiquer ni les raisons d'une réussite ou d'un échec, ni le potentiel d'apprentissage ultérieur du sujet. Aussi allons-nous voir plus bas que les applications essentielles du TAO qui ont été réalisées jusqu'à aujourd'hui se situent plutôt dans le domaine de l'évaluation sommative et certificative. La troisième génération de tests informatisés décrite par Bunderson et al. (1989) est appelée mesure continue *(continuous measurement, CM)*. Son fonctionnement technique est décrit comme étant proche du TAO, mais l'évaluation est censée être délivrée en continu, de manière à être intégrée dans le cursus de formation et dans les activités d'apprentissage avec l'objectif explicite de guider le processus d'apprentissage. Le résultat de la mesure devient ainsi dynamique et on prend ici résolument l'optique d'une évaluation formative. La quatrième génération de tests informatisés va encore aller un pas plus loin dans cette direction en essayant d'automatiser les mesures formatives à adopter à la suite des évaluations continues qui sont réalisées. Dans cette quatrième génération, le système informatique devrait permettre de garder continuellement à jour un modèle de l'état cognitif actuel de l'apprenant qui sera comparé à un état cognitif souhaité. L'ordinateur réalise ainsi un diagnostic permettant de choisir les mesures d'instruction adaptées à l'apprenant dans une base de mesures pédagogiques préalablement mise au point. On est donc très proche, ici, du fonctionnement d'un tuteur intelligent (Orey & Nelson, 1994) et cette quatrième génération est d'ailleurs qualifiée de mesure intelligente (intelligent measurement, IM). Remarquons encore que Bunderson et al. prévoyaient déjà en 1989 des applications concrètes pour les deux premières générations de tests informatisés, mais qu'ils

Psychologie et Psychométrie. Vol. 24 - N° 2/3 - 2003

n'étaient pas capables de présenter des exemples pleinement opérationnels ni pour la mesure continue, ni pour la mesure intelligente. Comme nous allons le voir dans la suite, cet état des choses n'a pas fondamentalement changé jusqu'à aujourd'hui.

Le TAO dans l'évaluation sommative et certificative

Dès la mise au point théorique du TAO, les avantages de cette forme d'informatisation des tests ont semblé évidents. Non seulement, on s'attendait à une individualisation du testing allant de pair avec une efficacité accrue de la passation (une meilleure précision de la mesure avec moins d'items), mais de plus on voyait encore d'autres avantages dans la passation informatisée, comme une sécurité plus élevée du test, une possibilité d'évaluation automatique et immédiate et celle de développer de nouveaux formats d'items (Green, 1983). Ces perspectives alléchantes ont eu pour effet qu'au courant des années 90, on a vu (notamment aux Etats-Unis) qu'un certain nombre de programmes d'évaluation à grande échelle ont été transposés du format papier-crayon classique à un format TAO. Parmi ces programmes de testing américains on trouve ainsi le *Graduate Record Examination, le Computerized Placement Test*, le *Test of English as a Foreign Language* ou le programme de certification du *National Council of State Boards of Nursing* (Zara, 1999) et du *National Board of Medical Examiners* (Luecht, 1996). En dehors des Etats-Unis et notamment en Europe, les programmes de testing basés sur le TAO sont beaucoup plus rares, mais commencent néanmoins à émerger, comme par exemple aux Pays-Bas où l'Institut National de la Mesure en Education vient de mettre au point un programme de TAO pour assigner des apprenants à différents niveaux de formation en mathématiques (cf. Eggen & Straetmans, 2000), ainsi que pour évaluer les performances dans des cours de mathématiques. Les contenus de ces programmes d'évaluation montrent donc qu'ils se situent principalement dans un contexte éducatif et que l'optique est essentiellement celle d'une évaluation sommative et certificative. Si on analyse les résultats obtenus par ces programmes (surtout les programmes américains qui fonctionnent depuis plusieurs années), on doit constater qu'ils ont connu un succès indéniable en ce qui concerne le volume des passations réalisées. Ainsi Wainer (2000a) indique un accroissement exponentiel dans la passation de TAO depuis le début des années 90 jusqu'à aujourd'hui.

Le problème de la sécurité des épreuves

La mise en pratique du TAO à grande échelle a également montré que toutes les attentes qu'on pouvait initialement avoir vis-à-vis du TAO n'ont pas pu être réalisées. Un problème majeur qui est apparu et qui se pose notamment pour des évaluations certificatives présentant un enjeu important pour les sujets, est celui

de la sécurité des épreuves. Il s'est en effet avéré que suite à l'algorithme de choix des items (le plus souvent l'algorithme du maximum d'information, cf. plus haut), on constate que la probabilité qu'un item spécifique soit choisi lors d'une passation donnée varie considérablement en fonction de la difficulté des items. Ainsi certains items sont très souvent sélectionnés au point qu'entre 15 et 20% des items constituant une banque d'items représentent plus de 50% des items qui sont administrés lors des passations réelles, alors que d'autres items ne sont que très rarement administrés (Wainer, 2000a). Ceci a pour conséquence que le «vol d'items» par le biais d'une mémorisation systématique d'un certain nombre d'items par des sujets passant successivement le même test peut sérieusement compromettre la sécurité de l'épreuve. Ce danger est d'autant plus grand que les programmes certificatifs de TAO à grande échelle sont obligés d'offrir des passations en continu, puisque des passations en grands groupes offertes à un nombre limité de dates au courant d'une année (ce qui était pratique courante pour les épreuves papier-crayon) ne sont pas envisageables pour le TAO suite à une disponibilité limitée de postes de travail informatisés. En fait il existe une relation linéaire entre le logarithme de la fréquence d'apparition d'un item dans un test et son rang d'utilisation (Zipf, 1949), c'est-à-dire qu'en augmentant de manière linéaire le rang d'utilisation d'un item, sa fréquence d'apparition dans un test va augmenter de manière exponentielle. Ceci a pour conséquence que la taille opératoire de la banque d'items est nettement moins importante que sa taille réelle. Wainer (2000b) arrive ainsi à la conclusion que pour augmenter la sécurité d'un TAO d'une manière linéaire, on est en fait obligé d'augmenter la taille de la banque d'items d'une manière exponentielle, ce qui est évidemment très difficilement réalisable en pratique. Il propose alors plusieurs stratégies de changement de l'algorithme de sélection des items, mais qui ont toutes pour conséquence d'abandonner le principe du choix des items montrant une adaptation optimale au niveau de compétence du sujet, ce qui implique donc la perte d'un des avantages principaux du TAO (Wainer, 2000b).

Une exploitation insuffisante des possibilités offertes par l'ordinateur

Une autre attente qu'on pouvait avoir vis-à-vis du TAO était la mise en œuvre de nouveaux formats d'items suite aux possibilités d'affichage et de retour multisensoriel qui sont offertes par l'ordinateur et qui ne sont pas disponibles pour des passations sous format papier-crayon. On pouvait ainsi envisager des items à caractère plus interactif utilisant des éléments graphiques animés, des simulations, des éléments sonores ou encore des séquences d'affichage basées sur des données chronométriques très précises. Or sous ce point de vue aussi, les programmes de TAO à grande échelle qui ont vu le jour jusqu'à aujourd'hui sont plutôt décevants, puisque les formats d'item qui sont utilisés sont le plus souvent des formats qui se prêteraient tout aussi bien à des passations papier-crayon. Ceci

Psychologie et Psychométrie. Vol. 24 - N° 2/3 - 2003

s'explique par le fait que les programmes en question sont pour la plupart issus d'outils d'évaluation préexistants et on a naturellement essayé de valoriser des banques d'item déjà constituées à partir de tests papier-crayon en les transposant sur support informatisé. Il faut en plus remarquer que l'utilisation principale des MRI pour la constitution des banques d'items utilisées dans le cadre du TAO impose des contraintes très fortes sur le nombre de sujets qui sont nécessaires pour étalonner de nouveaux items (plus de 1000 sujets pour des MRI à trois paramètres) et il est donc souvent plus facile et moins coûteux de réaliser des passations papier-crayon plutôt que des passations informatisées à des fins d'étalonnage. Ceci est d'autant plus vrai que des recherches ont pu montrer la stabilité des paramètres d'item lorsqu'on les transpose d'un support papier-crayon vers un support informatique (Mead & Drasgow, 1993; Olsen, Maynes, Slawson & Ho, 1989). Or une pratique d'étalonnage au format papier-crayon va empêcher l'exploitation des possibilités d'affichage qui sont offertes par l'ordinateur.

Un bilan plutôt décevant

Si on compare les résultats obtenus par les programmes de TAO à visée certificative aux attentes initiales qui étaient formulées vis-à-vis du TAO, on doit constater que beaucoup n'ont pu être réalisées et que la complexité de mise au point et de maintenance d'un programme de TAO à grande échelle s'est avérée plus importante qu'initialement prévue (Wise & Kingsbury, 2000). Au vu de ces résultats, Wainer (2000a) arrive à la conclusion que la mise en place d'un TAO ne devrait se faire que dans une situation où ce dispositif particulier puisse amener des avantages pour le problème d'évaluation particulier qui est posé. Dans le cas d'une évaluation certificative à grande échelle et à enjeu important, le TAO ne semble pas particulièrement conseillé, parce qu'il implique des problèmes de sécurité importants qui sont accentués par la nécessité de réaliser des passations continues dans le temps qui signifient une exposition continue de la banque d'items. Si en plus ces évaluations certificatives reposent essentiellement sur des items dont le format se prêterait également à des passations papier-crayon, l'utilité d'une transposition sur TAO peut fortement être mise en doute. Si on utilise le TAO par contre dans une optique formative et non sommative, il est clair que la question de la sécurité ne se pose plus, puisque dans ce cas, le sujet qui se trouve évalué a intérêt à ce que le résultat de l'évaluation reflète son véritable niveau de compétence, comme ce n'est qu'à cette condition qu'il va pouvoir recevoir une aide pédagogique subséquente la plus adaptée à son niveau de connaissances actuel. De même, dans cette optique formative, le caractère continu de l'évaluation sera souhaitable (au sens de la mesure continue postulée par Bunderson et al. 1989) et non plus une conséquence gênante, mais inévitable, du dispositif technique. Si on ajoute à cela la mise en œuvre de nouveaux formats d'items, on voit que c'est dans la mise en œuvre du TAO en tant qu'outil

d'évaluation formative qu'on peut trouver son plus grand potentiel de valorisation dans le domaine de l'éducation. Nous allons traiter ce point dans la suite.

Le TAO comme outil d'évaluation formative dans l'apprentissage de maîtrise (mastery learning)

La pédagogie de maîtrise est introduite par Bloom (1973, 1979) qui considère que la cause principale de l'échec dans les apprentissages scolaires tient au fait qu'une proportion importante d'élèves abordent une unité d'apprentissage nouvelle sans avoir maîtrisé les notions de l'unité précédente, qui constituent pourtant des prérequis pour l'unité nouvelle. Il propose donc d'évaluer les connaissances acquises à la fin de chaque unité, de remédier aux lacunes constatées et de n'aborder l'unité nouvelle que lorsque l'élève a maîtrisé les notions requises pour le nouvel apprentissage. En tant qu'approche évaluative de la pédagogie de maîtrise, Bloom, Hastings et Madaus (1971) introduisent le concept d'évaluation formative, les fonctions formative et sommative de l'évaluation ayant auparavant été introduites par Scriven (1967) dans le cadre de l'évaluation de programmes d'étude (cf. à ce sujet également Rouiller, 1998). Il en découle une des applications du TAO dans un contexte éducatif, à savoir l'apprentissage de maîtrise *(mastery learning)* dont la procédure d'évaluation - le testing de maîtrise *(mastery testing)* - peut être transposée sous forme de test adaptatif, ce qui implique un système d'évaluation qui est généralement désigné comme testing de maîtrise adaptatif (*adaptive mastery testing*, cf. Kingsbury & Weiss, 1983 ;Weiss & Kingsbury, 1984). Le principe de l'apprentissage de maîtrise et du testing de maîtrise (en dehors de toute considération de TAO) est assez simple (cf. Guskey, 1987). Il s'agit en fait de déterminer un seuil de performance au-delà duquel on va considérer qu'il y a maîtrise des apprentissages visés et en-dessous duquel on considèrera qu'il y a non-maîtrise de ces mêmes apprentissages. Ce seuil de performance séparant le niveau de maîtrise de celui de non-maîtrise est en général rendu opérationnel par un taux de réussite déterminé (en général assez élevé) sur les items d'un test couvrant le contenu d'apprentissage en question (on va par exemple considérer qu'une matière est maîtrisée par un sujet s'il arrive à répondre correctement à au moins 85% des questions d'un test couvrant cette matière). Le test en question est appelé test de maîtrise et le principe de l'apprentissage de maîtrise consiste dans le fait de faire revenir les apprenants sur l'étude d'une même matière aussi longtemps qu'ils n'ont pas atteint le niveau de maîtrise dans cette matière. En général, la matière est organisée par modules hiérarchiques et c'est seulement après avoir acquis le niveau de maîtrise pour un module donné qu'on fait avancer les apprenants vers le module suivant. L'évaluation qui est réalisée par les tests de maîtrise peut donc en partie être considérée comme sommative, puisqu'elle va en quelque sorte se

Psychologie et Psychométrie. Vol. 24 - N° 2/3 - 2003

contenter de certifier l'acquisition d'un certain degré de maîtrise, mais elle a également une très nette visée formative, puisque l'objectif primordial de l'évaluation est le feed-back pour l'étudiant. En cas de non-maîtrise, celui-ci va retravailler la matière sur laquelle ses apprentissages ont porté (en allant éventuellement briguer de l'aide auprès d'un enseignant), alors que dans le cas contraire il va avancer vers une matière de complexité plus élevée. L'efficacité de cette approche pédagogique a d'ailleurs été montrée dans plusieurs travaux (Bloom, 1973 ; Crahay, 2000 ; Guskey & Pigott, 1988 ; Kulik, Kulik & Bangert Drowns, 1990 ; Liefeld & Herrmann, 1990).

Lorsqu'on considère plus spécifiquement le problème de mesure qui est posé dans le cadre de l'apprentissage de maîtrise, on constate que l'objectif du test de maîtrise est en fait de réaliser une classification. On n'a donc pas forcément besoin de placer le sujet d'une manière précise sur un continuum latent (tel que cela est normalement réalisé dans le TAO, cf. plus haut), mais il faut qu'on puisse décider si le sujet se situe au-dessus ou en-dessous du niveau de maîtrise, c'est-à-dire s'il se situe dans le groupe des maîtres *(masters)* ou des non-maîtres (*non-masters*). Or on a constaté qu'en prenant un taux de réussite à un ensemble prédéterminé d'items comme critère d'attribution à ces deux groupes, on risque de commettre entre 40 et 60% d'erreurs de classification si les sujets sont proches du seuil de maîtrise et qu'on utilise des tests relativement courts (cf. Frick, 1990; Welch & Frick, 1993). En effet, la fidélité limitée de tels tests implique des intervalles de confiance assez étendus pour les scores vrais des sujets, ce qui explique le grand nombre de fausses classifications dans ce cas de figure. Les différentes procédures alternatives qui ont été proposées pour arriver à une prise de décision plus fiable dans ce problème de catégorisation seront présentées dans la suite.

Le testing de maîtrise adaptatif basé sur l'estimation statistique du niveau de compétence dans le cadre de la théorie de réponse à l'item

Weiss et Kingsbury (1984) ont proposé une procédure de TAO dans le cadre de l'apprentissage de maîtrise qui est directement basée sur l'estimation du niveau de compétence θ d'un sujet sur le continuum latent défini dans le cadre de l'application des MRI. On se situe donc ici dans le cadre général de fonctionnement du TAO qui a été décrit plus haut. Le problème spécifique qui se pose cependant dans ce cas de figure est celui de déterminer si le niveau de compétence estimé d'un sujet se trouve au-dessus ou en-dessous d'un seuil de maîtrise θ_m. Un premier problème qui se pose alors est la détermination du niveau θ_m qui se trouve sur la dimension latente définie par les MRI et dont l'origine est arbitraire, alors que les seuils qui sont en général appliqués dans le cadre de l'apprentissage de maîtrise s'expriment normalement sous forme d'un pourcentage d'items réussis sur en ensemble total d'items couvrant le domaine d'appren-

tissage. Mais la fonction de réponse du test (que l'on obtient en additionnant les différentes fonctions de réponse aux items, addition permise par la propriété d'indépendance locale des items; pour plus de détails, cf. par exemple Hambleton, Swaminathan & Rogers, 1991) indique pour chaque niveau θ le nombre de réponses correctes (et donc également le pourcentage de réponses correctes) qu'on devrait observer pour un sujet situé à ce niveau et ayant passé l'ensemble des items. Cette fonction permet alors, par une lecture inverse, de déterminer, pour un pourcentage de réponses correctes donné, quel est le niveau de compétence auquel ce pourcentage est normalement obtenu, ce qui permet donc de déterminer le niveau θ_m recherché. Une fois le niveau θ_m fixé, le deuxième problème qu'il reste à résoudre est de déterminer à partir de quand une décision de maîtrise versus non-maîtrise pourra être prise, ce qui permettra de fixer un critère d'arrêt pour la procédure de testing qui sera spécifiquement adapté au type de prise de décision exigé dans le cadre du test de maîtrise. Or la plupart des algorithmes d'estimation de la compétence, que ce soit l'algorithme du maximum de vraisemblance ou la méthode bayesienne de Owen (1975), permettent de calculer la variance d'erreur pour les niveaux de compétence estimés, ce qui permet donc la détermination de leurs intervalles de confiance. A partir de cette considération, il est assez facile d'arriver à une prise de décision au sens de la maîtrise, puisque l'administration de nouveaux items va continuer aussi longtemps que l'intervalle de confiance à 95% pour le niveau de compétence va contenir le seuil de maîtrise θ_m. En effet, tant que θ_m fait partie de l'intervalle de confiance en question, on peut considérer qu'il est impossible de juger si le niveau de compétence réel du sujet est situé au-dessus ou en-dessous de θ_m. Or dès que l'intervalle de confiance pour θ se trouve entièrement au-dessus ou en-dessous de θ_m une telle prise de décision est possible, avec néanmoins un risque de 2.5% (dans le cas d'un intervalle de confiance de 95%) de prendre la mauvaise décision. C'est donc à partir de l'étendue de l'intervalle de confiance adopté qu'on pourra varier le risque qu'on juge acceptable pour la prise de décision. La procédure de TAO telle qu'elle vient d'être décrite va impliquer que des sujets dont le niveau de compétence réel sera plus proche de θ_m vont se voir administrer des tests plus longs avant d'arriver à une prise de décision que des sujets dont le niveau de compétence réel sera plus éloigné de θ_m. Remarquons que la procédure de TAO qui vient d'être décrite peut facilement être étendue à des problèmes de catégorisation comportant plus de deux groupes (au lieu de faire la différenciation entre un groupe de non-maîtrise et un groupe de maîtrise, on peut envisager une classification dans des groupes de niveau). Weiss et Kingsbury (1984) présentent d'ailleurs un modèle pour l'assignation adaptative de niveaux de performance (*adaptive grading test*) qui repose sur cette procédure. Ce qui change pour le cas d'une classification en N groupes est le fait qu'il faudra prévoir non plus un seul seuil décisionnel θ_m, mais N-1 seuils de ce type. La décision de classification dans un niveau sera alors prise à partir du moment où l'intervalle de confiance pour θ sera entièrement contenu entre deux seuils successifs délimitant ce niveau de performance. Des variations de cette procédure de classification sont présentées par Xiao (1999).

Psychologie et Psychométrie. Vol. 24 - N° 2/3 - 2003

La critique que certains auteurs formulent vis-à-vis de cette procédure de testing de maîtrise adaptatif est qu'elle repose sur une procédure d'estimation assez complexe qui vise le positionnement du sujet sur un continuum, alors qu'on se trouve en présence d'une prise de décision catégorielle ordinale (voir même nominale) qui pourrait éventuellement reposer sur des procédures de décision plus simples (Eggen & Straetmans, 2000). Une deuxième critique qui est reliée à la première provient du fait que la procédure décrite repose sur la mise en œuvre de la théorie de réponse à l'item, avec toutes les contraintes que ce modèle de mesure comporte pour l'étalonnage des items limitant sa mise en œuvre dans un contexte scolaire quotidien (Welch & Frick, 1993). Aussi a-t-on essayé de proposer des procédures alternatives qui soient mathématiquement plus simples pour réaliser du TAO dans l'optique d'un apprentissage de maîtrise. Ces procédures seront présentées dans la suite.

Le testing de maîtrise adaptatif basé sur le test séquentiel du rapport de probabilité (sequential probability ratio test, SPRT)

La procédure alternative la plus utilisée et la plus étudiée repose sur le test séquentiel du rapport de probabilité développé par Wald (1947). Wald a pu montrer que par rapport à une méthode traditionnelle qui consiste à déterminer a priori un ensemble d'observations qui sont à réaliser pour ensuite appliquer un test statistique à l'ensemble de ces observations, il est nettement plus avantageux et efficace de réaliser ces observations d'une manière séquentielle et d'appliquer un certain nombre de lois décisionnelles après chacune de ces observations. Une telle procédure séquentielle permet en effet de réduire de moitié le nombre d'observations nécessaires sans pour autant augmenter le taux d'erreurs de type I et II dans les décisions de classification qui sont à prendre. Un certain nombre de travaux ont pu montrer qu'il est possible d'utiliser le modèle SPRT en tant que modèle décisionnel pour le testing de maîtrise adaptatif où il s'agit donc de collecter des informations sous forme de réponses à des items qui doivent permettre d'assigner les sujets soit au groupe de maîtrise, soit à celui de non-maîtrise (Frick, 1989 ; 1990 ; Lewis & Sheehan, 1990 ; Reckase, 1983). Sur la base du nombre de réponses justes (j) et du nombre de réponses fausses (f) et en connaissant le taux de réponses correctes dans le groupe de maîtrise (P_m), ainsi que celui dans le groupe de non-maîtrise (P_n), le rapport de probabilité PR peut être calculé :

$$PR = \frac{P_m^j(1 - P_m)^y}{P_n^j(1 - P_n)^y}$$

Sur la base des trois règles décisionnelles suivantes, il est alors possible de déterminer pour chaque sujet la longueur du test qu'il faut lui administrer, afin de pouvoir arriver à une décision quant à son appartenance de groupe :

- si

$PR \geq (1 - \beta) / \alpha$, alors choisir l'hypothèse de maîtrise et arrêter les observations

- si

$PR \leq \beta / (1 - \alpha)$, alors choisir l'hypothèse de non-maîtrise et arrêter les observations

- si

$\beta / (1 - \alpha) \leq PR \leq (1 - \beta) / \alpha$, alors choisir un autre item et continuer les observations

α représente le risque d'attribuer une personne au groupe de maîtrise alors qu'elle appartient au groupe de non-maîtrise et β le risque d'attribuer une personne au groupe de non-maîtrise, alors qu'elle appartient au groupe de maîtrise. C'est donc à partir de ces deux paramètres qu'on peut varier le risque d'erreur décisionnelle qui est jugé acceptable.

Cette procédure peut être appliquée en se basant sur un ensemble d'items qui ont été étalonnés à l'aide des MRI (Lewis & Sheehan, 1990). Sur une telle banque d'items, on va alors utiliser des algorithmes de sélection des items simples qui sont adaptés au modèle *SPRT*. En général, on choisira les items qui vont fournir un maximum d'information au niveau de compétence θ_m correspondant au seuil de maîtrise. Certains travaux ont d'ailleurs pu montrer que pour des décisions catégorielles, cette stratégie de choix des items s'avère aussi efficace que l'algorithme du maximum d'information au niveau de compétence estimé (Frick, 1990). Avec le modèle *SPRT*, on va donc perdre dans le TAO le volet adaptatif qui consistait à adapter le degré de difficulté des items au niveau de compétence du sujet, mais on va garder une adaptation de la longueur du test qui est requise pour arriver à une décision au sens de maîtrise *versus* non-maîtrise. Les sujets dont la compétence est plus proche du seuil de maîtrise se verront administrer des tests plus longs que ceux dont la compétence en est plus éloignée. Le grand attrait du modèle *SPRT* est sa simplicité mathématique qui s'avère pourtant efficace, puisque sur des problèmes décisionnels catégoriels, le modèle semble produire le même taux de classements corrects que le testing de maîtrise adaptatif basé sur les MRI, en utilisant un nombre d'items qui peut même être inférieur (cf. plus bas).

Psychologie et Psychométrie. Vol. 24 - N° 2/3 - 2003

Un autre avantage du modèle *SPRT* est le fait qu'il peut être appliqué sans forcément recourir aux MRI. Cet aspect est souligné par Welch et Frick (1993) qui remarquent que les contraintes très fortes pour étalonner les items à l'aide des MRI (notamment en ce qui concerne la taille des échantillons), ainsi que leur complexité mathématique, limitent leur utilisation par des enseignants dans des environnements d'apprentissage «naturels». Pour une mise en œuvre efficace du modèle *SPRT*, il suffit par contre de disposer de données d'étalonnage recueillies sur un échantillon représentatif d'environ 50 sujets, en veillant néanmoins à ce que cet échantillon soit composé d'une moitié de sujets ayant un niveau de maîtrise et d'une moitié de sujets ayant un niveau de non-maîtrise. Le problème qui existe avec le modèle *SPRT* est que, dans sa version initiale, il ne tient pas compte des paramètres d'items (notamment de la difficulté) et qu'il traite donc tous les items comme étant équivalents. Cela implique que le choix des items se fait au hasard lorsque le modèle décisionnel arrive à la conclusion que la collecte des observations doit continuer. Ceci constitue donc un désavantage majeur par rapport aux algorithmes d'optimisation du choix des items qui peuvent être appliqués dans le cadre des MRI. Or malgré cette limite évidente, plusieurs travaux ont pu montrer que l'efficacité décisionnelle du modèle *SPRT* est équivalente à celle obtenue avec des tests de maîtrise adaptatifs basés sur les MRI (Eggen & Straetmans, 2000 ; Frick, 1990 ; 1992 ; Welch & Frick, 1993). Frick (1992) a d'ailleurs proposé deux extensions du modèle *SPRT* qui sont censées tenir compte de la perte d'information engendrée par le choix aléatoire des items. La première extension (dénommée *EXSPRT*) combine le modèle *SPRT* avec des possibilités de raisonnement du type système expert. Cette extension implique que les items sont toujours choisis au hasard, mais le système expert leur assigne des poids différents en fonction de la probabilité de réussite différentielle des items dans le groupe de maîtrise et de non-maîtrise. La deuxième extension (denommée *EXSPRT*-I) introduit une méthode «intelligente» de choix des items qui va donc permettre de réintroduire un caractère adaptatif faisant en sorte que la difficulté des items sera adaptée à la compétence du sujet. Les possibilités de mise en pratique de *EXSPRT* et de *EXSPRT*-I ont été montrées par Welch et Frick (1993). Remarquons encore qu'en combinant plusieurs modèles *SPRT*, on arrive à des généralisations du modèle pour réaliser une classification dans trois ou même dans K catégories (cf. Eggen & Straetmans, 2000). On peut donc retenir que le modèle *SPRT* et ses modèles dérivés peuvent constituer une alternative intéressante aux algorithmes de testing adaptatif basés sur les MRI dans le cas où on se trouve confronté à une prise de décision catégorielle. Lorsqu'il s'agit par contre de réaliser une estimation précise de la position d'un sujet sur le continuum latent, le testing adaptatif réalisé dans le cadre des MRI paraît plus performant (Frick, 1990; Kingsbury & Weiss, 1983).

Le TAO comme outil d'évaluation formative dans l'évaluation diagnostique visant à guider les interventions pédagogiques

Même si le testing de maîtrise comporte déjà un élément très fort d'évaluation formative puisque l'objectif pédagogique est de donner un retour aux apprenants qui leur permette de remédier à d'éventuelles lacunes, on doit néanmoins constater que le diagnostic de maîtrise versus non-maîtrise est en soi un diagnostic sommatif sous forme de constat statique qui ne permet pas de guider d'une manière plus précise la boucle de remédiation subséquente. Or l'objectif d'une mesure continue, et plus encore d'une mesure intelligente, telles qu'elles étaient présentées par Bunderson et al. (1989), est de fournir des informations précises concernant le fonctionnement cognitif actuel de l'apprenant qui puissent être utilisées pour mieux cibler les interventions pédagogiques qu'on va lui proposer, afin qu'il puisse progresser au mieux à partir de son état cognitif actuel (cf. à ce sujet Grégoire, 1996). Une telle évaluation diagnostique n'est certainement pas atteinte par le testing de maîtrise adaptatif tel qu'il est présenté plus haut. Une procédure de TAO qui approche déjà plus le concept de mesure continue est exposée par Weiss et Kingsbury (1984) qui proposent de réaliser à l'aide d'un TAO des mesures successives du niveau de compétence des sujets au cours d'un processus d'apprentissage (ils parlent d'une «mesure adaptative des changements individuels de performance» - *adaptive measurement of individual changes in achievement*). Or même si on arrive à situer un sujet au cours d'un apprentissage à des positions successives précises sur la dimension latente qui est définie par les MRI, un tel positionnement permet bien de retracer un progrès d'apprentissage sur un continuum, mais ne permet pas de diagnostiquer d'une manière précise le fonctionnement cognitif d'un sujet à un moment donné qui puisse alors être relié d'une manière théoriquement fondée à des interventions pédagogiques particulièrement adéquates. Afin d'approcher cet objectif d'un diagnostic plus précis, il paraît indispensable d'établir des liens plus étroits entre les procédures de TAO actuellement opérationnelles et des développements récents dans le domaine de la psychologie cognitive. Dans leur description des défis futurs auxquels le TAO devra faire face, Wainer et al. (2000) soulignent d'ailleurs que le TAO devrait s'efforcer de réaliser une intégration plus importante des résultats de la psychologie cognitive dans la mise au point des tests, notamment en vue d'une interprétation plus claire des résultats obtenus. Une telle approche de «psychométrie cognitive» visant l'intégration entre psychométrie et psychologie cognitive afin d'établir des diagnostics plus précis des processus de résolution de problèmes mis en œuvre par des sujets sur une tâche spécifique est recherchée depuis plusieurs années en dehors du domaine du TAO (cf. par exemple Beuscart-Zéphir, Anceaux, Duhamel & Quenton, 1996; Houssemand, 2001; Martin, 1999; Richard & Zamani, 1996). En TAO, il existe, à l'heure actuelle, différentes tentatives pour réaliser, au moins à l'échelle d'un

Psychologie et Psychométrie. Vol. 24 - N° 2/3 - 2003

environnement expérimental, une telle intégration de la psychologie cognitive. Plusieurs de ces travaux seront présentés dans la suite.

Le modèle multidimensionnel de Rasch pour l'apprentissage et le changement de Embretson (Multidimensional Rasch Model for Learning and Change, MRMLC)

L'objectif initial qui est poursuivi par le *Multidimensional Rasch Model for Learning and Change (MRMLC)* est de donner une réponse aux problèmes qui sont posés par la détermination de scores de gain cognitif calculés sous forme de différence entre scores bruts à un test obtenus après et avant un apprentissage. Ces scores sont en effet peu fiables, présentent une métrique douteuse et impliquent des corrélations négatives artefactuelles (Bereiter, 1963). La solution proposée par Embretson (1991) se présente sous forme d'un modèle de Rasch multidimensionnel. Celui-ci va estimer des aptitudes latentes multiples à partir d'items présentant un seul paramètre de difficulté, mais avec la particularité que ces aptitudes multiples vont représenter une aptitude initiale, ainsi qu'une ou plusieurs «modifiabilités» *(modifiabilities)* qui sont autant de paramètres de changement correspondant aux gains d'apprentissage que le sujet a réalisés entre deux moments de mesure successifs. Ainsi, pour un plan expérimental sous forme de prétest-intervention-posttest, avec deux moments de mesure (prétest et posttest), *MRMLC* va estimer deux paramètres de personne pour chaque sujet, à savoir un paramètre θ_1 qui correspondra au niveau de compétence initial (au moment du prétest) et un paramètre de modifiabilité θ_2 qui correspondra au changement intervenu entre le prétest et le posttest. Afin de résoudre l'ensemble des problèmes posés par les scores de gain classiques à l'aide de *MRMLC*, il est important de minimiser l'erreur de mesure à la fois pour le niveau de compétence initial θ_1 et pour la modifiabilité θ_2. Dans ce contexte, Embretson (1995) souligne que c'est justement par le biais du testing adaptatif qu'une telle minimisation de l'erreur pourra être atteinte. En effet, comme le *MRMLC* est un MRI, on pourra l'implémenter sous forme de TAO et choisir les items de manière à maximiser l'information fournie à la fois pour le niveau de compétence initial et pour la modifiabilité.

Dans le contexte de l'utilisation du TAO comme outil de diagnostic visant à guider les interventions pédagogiques, un intérêt particulier revient à une extension de *MRMLC* (désignée par *MRMLC+*) et dont l'objectif est précisément de faire le lien entre, d'un côté, les progrès d'apprentissage d'un sujet et, d'un autre côté, la nature des traitements cognitifs et des représentations qui sont impliqués dans ces progrès (Embretson, 1995). Pour atteindre cet objectif, on va remplacer le paramètre de difficulté dans le modèle *MRMLC+* par une combinaison linéaire de paramètres de complexité caractérisant l'item et qui seront autant de

prédicteurs pour la difficulté de celui-ci. Pour des problèmes mathématiques présentés sous forme verbale, il s'avère par exemple que le nombre de mots utilisés pour la présentation du problème est un bon prédicteur de la difficulté de l'item (Embretson, 1995). Dans *MRMLC+*, la difficulté de l'item va donc être estimée à partir d'une combinaison linéaire de plusieurs paramètres structuraux de ce type définissant autant d'aspects différents de la complexité des traitements cognitifs requis pour la résolution de l'item. *MRMLC+* permettra ainsi de placer sur une même échelle de compétence latente à la fois la compétence initiale des sujets, la modifiabilité exprimant le gain d'apprentissage et la difficulté des items qui se laissera expliquer par les paramètres structuraux retenus pour définir la complexité des traitements cognitifs requis. Ce dernier point est essentiel dans l'utilisation du *MRMLC+* en tant qu'outil de diagnostic destiné à guider des interventions pédagogiques subséquentes, puisque le changement de compétence observé ne pourra pas seulement être interprété comme avancement quantitatif sur la dimension latente. Au-delà de cette interprétation quantitative, le modèle prédictif pour la difficulté des items permettra une interprétation en termes des paramètres de complexité qui sont maîtrisés à la suite d'un certain apprentissage et d'identifier également les paramètres de complexité qui ne sont pas encore maîtrisés, alors qu'ils constituent l'étape suivante dans la progression de difficulté des items. Un même changement en termes de quantité pourra alors recevoir des interprétations qualitativement différentes en fonction de la compétence initiale du sujet et en fonction des paramètres de complexité nouvellement maîtrisés, ce qui permettra également de tirer des conclusions plus individualisées en ce qui concerne les conséquences instructives. La viabilité et l'intérêt de *MRMLC+* ont été montrés par Embretson (1995) à l'exemple du raisonnement mathématique. On peut néanmoins remarquer que les décisions pédagogiques à prendre à la suite d'une évaluation par *MRMLC+* restent à charge de l'enseignant et que le modèle, même s'il facilite largement une démarche de diagnostic remédiatif, ne prévoit pas actuellement une automatisation du processus décisionnel concernant les interventions pédagogiques appropriées.

La méthode de l'espace-règles de Tatsuoka et Tatsuoka (rule-space method)

Tatsuoka et Tatsuoka (1997) proposent un système de TAO dont l'objectif n'est pas prioritairement la mesure d'un changement, mais qui se propose de réaliser le diagnostic d'un état de connaissances (*knowledge state*) actuel d'un apprenant, afin de pouvoir identifier d'une manière précise les apprentissages subséquents qui devront être réalisés par cet apprenant. Le système pourra dans la suite être utilisé pour diagnostiquer le succès de l'apprentissage, mais ce succès ne se traduira pas comme une progression quantitative sur une dimension latente

sous forme de paramètre de changement ou modifiabilité (comme chez Embretson), mais comme le passage d'un premier état de connaissances vers un état de connaissances suivant, généralement plus évolué.

Afin de pouvoir déterminer l'état de connaissances d'un apprenant, il faut définir a priori les états de connaissances possibles pour le domaine de connaissances visé. A cet effet, il est nécessaire d'identifier l'ensemble des processus cognitifs élémentaires qui sont requis pour résoudre les différents items couvrant le domaine d'apprentissage en question. Ces processus cognitifs élémentaires sont appelés des «attributs cognitifs» et la nature exacte d'un item donné peut alors être décrite à l'aide des attributs cognitifs que cet item va mettre en jeu (un item va en général impliquer uniquement un sous-ensemble des attributs cognitifs constituant le domaine de connaissances qui fait l'objet de l'apprentissage). Par exemple, pour des items concernant des problèmes mathématiques d'addition de fractions, le processus cognitif permettant de trouver le dénominateur commun est considéré comme un attribut cognitif spécifique décrivant ce type d'items (Tatsuoka & Tatsuoka, 1997). Un état de connaissances correspond alors à un ensemble d'attributs cognitifs qui sont maîtrisés par un sujet et on suppose qu'un item peut seulement être résolu par un sujet s'il maîtrise tous les attributs cognitifs qui sont impliqués dans son processus de résolution. Si on suppose qu'un domaine d'apprentissage implique K attributs cognitifs, on devrait théoriquement envisager l'existence de 2^K états de connaissances. Or, en pratique, une analyse de la fréquence des conceptions erronnées qu'on trouve pour un domaine de connaissances particulier montre que le nombre d'états de connaissances qu'on rencontre effectivement est nettement plus faible. L'objectif de la procédure de testing adaptatif diagnostique sera alors de déterminer aussi efficacement que possible l'état de connaissances dans lequel un sujet se trouve au moment de l'évaluation.

A cet effet, une procédure de testing adaptatif sera utilisée qui peut être considérée comme une extension de la méthode classiquement utilisée pour les items étalonnés dans le cadre des MRI (cf. plus haut). A partir des réponses qui sont données par le sujet, on va estimer son niveau de compétence θ, mais on va en plus estimer un paramètre qui correspondra au degré de typicalité ζ du patron de réponses qui a été donné par le sujet. L'espace cartésien bi-dimensionnel qui est ouvert par les deux dimensions latentes θ et ζ est appelé l'espace-règles (rule space, cf. Tatsuoka, 1983) et la procédure adaptative consiste, en fait, dans un déplacement à l'intérieur de cet espace-règles. On peut dès lors montrer qu'un état de connaissances particulier correspond à une paire de coordonnées précises à l'intérieur de cet espace-règles, et ainsi la position finale qui sera attribuée à un sujet dans cet espace permettra également de déterminer son état de connaissances actuel (comme étant l'état de connaissances prédéfini qui est le

plus proche de la position finale du sujet dans l'espace-règles). L'algorithme de sélection des items employé peut être considéré comme analogue à la méthode de la plus forte pente *(steepest descent)* qui est souvent employée dans des problèmes d'optimisation. Il consiste à choisir comme item suivant celui qui va minimiser la distance entre la position actuelle du sujet et son état de connaissances final (Tatsuoka & Tatsuoka, 1997).

Un autre avantage de la méthode proposée par Tatsuoka et Tatsuoka (1997) est le fait qu'elle permet d'automatiser les mesures pédagogiques devant être adoptées à la suite du diagnostic d'un état de connaissances particulier. Il sera en effet possible de déterminer quel est l'état de connaissances plus évolué qui est le plus proche de l'état de connaissances actuellement acquis par un sujet. Les mesures instructives à adopter vont alors se centrer sur l'attribut cognitif qui fait la différence entre ces deux états de connaissance. On doit néanmoins remarquer qu'une telle focalisation sur des attributs cognitifs précis présuppose une certaine additivité linéaire entre ces attributs. Elle ignore largement des interactions possibles et même probables et ne correspond donc pas forcément au fonctionnement cognitif réel du sujet.

Tatsuoka et Tatsuoka (1997) ont réalisé une mise en pratique de la méthode d'évaluation diagnostique adaptative de l'espace-règles à l'exemple de l'apprentissage des fractions en mathématiques. Les résultats de ce travail ont pu montrer la fidélité des états de connaissances déterminés à l'aide de l'espace-règles. L'efficacité potentielle d'un système d'instruction basé sur cette méthode de TAO et comportant des cycles de prétest - assignation à des programmes de remédiation - activités de remédiation - posttest, a également pu être établie.

Perspectives d'avenir

La mise au point de banques d'items générés par ordinateur

Même si les travaux présentés dans la partie précédente restent encore à un état expérimental, ils s'approchent néanmoins déjà beaucoup plus des systèmes de mesure continue et de mesure intelligente qui ont été prévus par Bunderson et al. (1989). Or ces travaux montrent également très clairement qu'une utilisation efficace des systèmes de TAO pour le guidage des interventions pédagogiques dans un contexte scolaire présuppose que le choix et la caractérisation des items reposent sur une base théorique solide en ce qui concerne la description des processus cognitifs impliqués. Une utilisation du TAO pour réaliser des diagnostics processuels précis va donc mettre l'accent sur le degré d'élaboration de la théorie définitoire (Dickes, Tournois, Flieller & Kop, 1994) qui est à la base de la banque d'items du TAO. Un objectif à moyen, voir à long terme qui peut alors

Psychologie et Psychométrie. Vol. 24 - N° 2/3 - 2003

être envisagé est la mise au point de théories de traitement cognitif qui soient d'un degré de précision qui est tel qu'on peut prédire le degré de difficulté des items à partir des exigences de traitement qu'ils impliquent (Kyllonen, 1997 ; Wainer et al. 2000). Cela présuppose qu'on puisse quantifier exactement la contribution relative des différents aspects du traitement de l'information à la difficulté des items et qu'on arrive également à identifier d'une manière précise les interactions éventuelles qui existent entre différents paramètres structurels contribuant au degré de complexité de la tâche. Si toutefois une telle modélisation exhaustive réussit, elle va comporter des avantages multiples, puisqu'elle va donner accès à des diagnostics très précis des aptitudes processuelles des sujets (en se basant sur le type d'items que les sujets arrivent à traiter efficacement) et elle va également résoudre le problème de l'étalonnage de nouveaux items et de la taille de la banque d'items. En effet, si l'on disposait de modèles prédictifs performants, on n'aurait plus besoin de passer par l'étalonnage coûteux des items sur des populations réelles; on pourrait générer de nouveaux items à paramètres spécifiques à l'aide de l'ordinateur, en utilisant directement le modèle prédictif.

Utilisation de nouveaux formats d'item

Un aspect qui doit aussi encore être développé à l'avenir est celui d'une plus ample exploitation des possibilités d'affichage et de retour qui sont offertes par l'ordinateur. Comme nous l'avons déjà évoqué plus haut, les systèmes de TAO qui ont été mis en pratique jusqu'à aujourd'hui recourent avant tout à des types d'items dont la passation serait également possible sous forme papier-crayon. Or des travaux récents en sciences cognitives (et notamment en neuropsychologie cognitive, cf. par exemple Engle, Kane & Tuholski, 1999) montrent que les performances dans des tâches complexes de type résolution de problèmes telles qu'elles sont classiquement utilisées dans les évaluations scolaires, reposent sur des processus cognitifs plus élémentaires de type perceptif ou mnésique qui sont souvent difficilement mesurables avec des dispositifs papier-crayon. Ainsi par exemple des dispositifs d'évaluation de la mémoire de travail impliquent souvent des éléments dynamiques qui pourraient tirer pleinement profit des possibilités d'affichage dynamique qui sont offertes par l'ordinateur. Les possibilités d'affichage enrichies de l'ordinateur (et notamment sa potentialité multimédia) permettent également l'évaluation d'aptitudes qui étaient jusqu'ici largement négligées par la psychométrie, suite à la difficulté de mise en œuvre des dispositifs d'évaluation y afférents. Ainsi Vispoel (1999) présente par exemple un TAO pour l'évaluation de l'aptitude musicale.

Un autre type de tâche, encore peu exploré quant à son utilité psychométrique, est l'utilisation de simulations qu'il est possible de réaliser à l'aide de l'ordinateur et qui pourraient donner accès à des tâches écologiquement plus

valides pour l'évaluation d'un grand nombre de compétences, comme par exemple certaines compétences professionnelles. Ainsi Hanson, Bormann, Mogilka, Manning et Hedge (1999) présentent un dispositif de simulation pour contrôleurs de trafic aérien sur la base duquel ils génèrent des questions à choix multiple permettant d'évaluer l'aptitude des sujets à réagir d'une manière adéquate à des situations-problème qui se présentent dans ce domaine professionnel précis. Mais aussi l'évaluation de compétences plus directement liées à un certain type de fonctionnement cognitif, comme par exemple l'évaluation de la cognition spatiale, pourraient tirer bénéfice de l'utilisation de simulations informatisées, comme cela a par exemple été montré par Martin (1997) pour des déplacements à l'intérieur d'un espace virtuel tri-dimensionnel. Un autre aspect technique qui est offert par l'ordinateur et qui reste encore largement sous-exploité est l'utilisation des temps de réaction enregistrés lors de la résolution des items. Or comme cela a été souligné par Martin et Houssemand (2002), la signification des temps de réaction recueillis sur des items classiques reste encore largement indéterminée et nécessite encore des efforts théoriques majeurs. Néanmoins l'ordinateur permettrait aussi de créer des tâches spécifiques qui seraient conçues de manière à faciliter l'interprétation des temps de réaction recueillis et pour lesquelles le temps de réaction serait ainsi l'indicateur le plus important de la performance à la tâche (cf. Martin, 1999). Néanmoins il reste encore un travail conceptuel important à réaliser pour permettre la mise en pratique de tous ces nouveaux formats d'items sous forme de TAO.

Mise en réseau à l'aide d'internet

Un autre aspect du TAO dans un contexte scolaire est celui de l'utilisation des résultats recueillis par un tel système d'évaluation. En effet, surtout dans un contexte éducatif et dans l'optique d'une évaluation formative et continue, il paraît important de pouvoir suivre la progression des élèves dans les différents domaines d'apprentissage à travers l'ensemble du cursus scolaire. De même, lorsqu'un enseignant se trouve en face d'un nouveau groupe-classe, des données recueillies à l'aide d'un système de TAO pourraient fournir des informations plus précises et plus utiles pour la mise en route du processus d'apprentissage avec ces nouveaux élèves que cela n'est actuellement possible avec les notes du bulletin peu fiables qui sont en général utilisées à cette fin dans un contexte scolaire. Or, une telle utilisation du TAO implique un stockage centralisé des données recueillies localement lors des passations de TAO par les élèves. Cette centralisation systématique permettrait également de combiner des données recueillies pour des finalités diverses et de dresser un tableau plus complet du niveau de connaissances et du niveau de progression des élèves participant au programme de TAO. Or cette centralisation systématique des données est

Psychologie et Psychométrie. Vol. 24 - N° 2/3 - 2003

difficilement réalisable sans une mise en réseau des ordinateurs servant de plate-forme au TAO. Ainsi Kingsbury et Hauser (1999) décrivent-ils un système de TAO mis en place dans le système scolaire d'une région des Etats-Unis comme outil d'évaluation principal couvrant les différentes tâches évaluatives qui incombent dans ce contexte scolaire précis (diagnostic, certification, tests d'entrée pour des programmes de formation spécifiques, progrès d'apprentissage, pré- et posttest pour des interventions pédagogiques expérimentales). Si le bilan global dressé pour ce programme de TAO est très positif, les auteurs soulignent qu'un de ses problèmes majeurs est le manque d'efficacité dans la centralisation des résultats causé par une mise en réseau insuffisante de la plate-forme de testing. A l'avenir, on peut néanmoins imaginer le développement de systèmes de TAO intégrés dont la mise en réseau se réalise à travers internet, ce qui permettra une centralisation facile des données même dans des programmes de TAO à grande échelle. De tels systèmes devraient également réaliser une mise à disposition des outils d'évaluation à distance permettant ainsi une mise à jour facile des contenus, ce qui est également particulièrement important pour un système de TAO devant fonctionner dans un contexte scolaire quotidien.

Conclusions

Comme cela est le cas pour beaucoup de révolutions annoncées à un moment donné de l'histoire, celle prévue à la fin des années 80 pour la mesure en éducation à l'occasion de l'introduction de l'ordinateur et notamment du testing adaptatif par ordinateur, n'a pas vraiment eu lieu. Plus d'une décennie plus tard, on constate que l'ordinateur s'est effectivement introduit dans quasiment tous les domaines de la vie quotidienne, qu'il est aussi de plus en plus présent dans nos écoles, mais que le rôle joué par l'ordinateur en tant qu'outil d'évaluation dans un contexte scolaire reste encore assez marginal, du moins en Europe. Les mises en pratique du TAO qu'on a vu émerger au cours de la dernière décennie concernent avant tout des programmes d'évaluation à grande échelle existant depuis longtemps aux Etats-Unis, or il s'est avéré que ces programmes certificatifs à grande échelle doivent probablement compter parmi les applications les moins conseillées du TAO. Trop importants paraissent les problèmes de sécurité des tests impliqués notamment par la contrainte de testing continu que la plate-forme informatique impose. En plus, le fait que ces programmes d'évaluation ont pour la plupart maintenu les formats d'item classiques (bien connus des passations papier-crayon) implique qu'un des atouts majeurs de la passation informatisée n'a pas été utilisé. Or, ce bilan plutôt négatif ne doit pas faire oublier que le potentiel du TAO est bien réel et que, sous la forme d'un outil d'évaluation formative, il pourrait bien contribuer à une individualisation de l'apprentissage scolaire. Sous la forme d'un système intégré délivré par réseau, le TAO pourrait être l'outil de travail majeur d'un enseignant qui voit son rôle principal dans la gestion individualisée des

processus d'apprentissage de ses élèves et qui doit donc créer une offre de formation qui est adaptée à l'état de développement actuel des apprenants. Un tel enseignant a besoin d'un outil de diagnostic efficace qui puisse être administré à plusieurs moments de l'année et dont la mise en œuvre ne soit pas trop coûteuse (à la fois d'un point de vue infrastructurel et d'un point de vue temporel). Le TAO avec son caractère adaptatif pourrait être un tel outil. Si on arrive en plus à combiner les contenus évalués par le TAO avec des théories précises concernant le fonctionnement cognitif, l'évaluation issue du TAO devrait permettre de formuler des hypothèses précises concernant les activités de formation les plus adaptées à un apprenant particulier, ces activités de formation pouvant à nouveau (du moins en partie) être délivrées par l'outil informatique. Une mise en œuvre efficace du TAO dans les années à venir ne constitue donc pas seulement un défi majeur pour le domaine de la psychométrie, mais également et même avant tout pour celui de la psychologie cognitive. En effet, les progrès réalisés dans la mise au point de nouveaux outils psychométriques sont souvent largement dépendants d'avancées dans le domaine de la psychologie cognitive qui sont indispensables pour la mise en œuvre de ces outils et pour l'interprétation des résultats qui en ressortent (cf. les travaux de Embretson, 1995 et de Tatsuoka & Tatsuoka, 1997 présentés plus haut). Un problème crucial à résoudre dans cette optique sera dès lors celui de créer des systèmes offrant un bon ancrage théorique, mais qui soient néanmoins assez simples pour être appliqués à grande échelle par des enseignants de terrain qui ne sont ni des spécialistes de la mesure, ni de la psychologie cognitive.

BIBLIOGRAPHIE

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C. Harris (Ed.), *Problems in measuring change* (pp. 3-20). Madison : University of Wisconsin Press.

Beuscart-Zéphir, M. C., Anceaux, F., Duhamel, A., & Quenton, S. (1996). Un exemple d'application du diagnostic cognitif, *Psychologie Française*, 14(1), 65-76.

Bloom, B. S. (1973). Recent developments in mastery learning. *Educational Psychologist*, 10(2), 53-57.

Bloom, B.S. (1979). *Caractéristiques individuelles et apprentissages scolaires*. Paris : Nathan.

Bloom, B.S., Hastings, J., & Madaus, G. (1971). *Handbook on formative and summative evaluation of student learning*. New York : McGraw Hill.

Bunderson, V. C., Inouye, D. K., & Olsen, J. B. (1989). The four generations of computerized educational measurement. In R. L. Linn (Ed.), *Educational measurement : Third edition* (pp. 367-407). New York : Macmillan.

Crahay, M. (2000). *L'école peut-elle être juste et efficace ? De l'égalité des chances à l'égalité des acquis*. Bruxelles : De Boeck.

Psychologie et Psychométrie. Vol. 24 - N° 2/3 - 2003

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

Dickes, P., Tournois, J., Flieller, A., & Kop, J.-L. (1994). *La psychométrie*. Paris : Presses Universitaires de France.

Dickes, P. (1999). Modèles de réponse à l'item (MRI) et recherche en psychologie. *Psychologie et Psychométrie*, 20(2-3), 9-18.

Eggen, T. J. H. M., & Straetmans, G. J. J. M. (2000). Computerized adaptive testing for classifying examinees into three categories. *Educational and Psychological Measurement*, 60(5), 713-734.

Embretson, S. E. (1991). A multidimensional latent trait model for measuring learning and change. *Psychometrika*, 56(3), 495-515.

Embretson, S. E. (1995). A measurement model for linking individual learning to processes and knowledge : Application to mathematical reasoning. *Journal of Educational Measurement*, 32(3), 277-294.

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

Frick, T. W. (1989). Bayesian adaptation during computer-based tests and computer-guided practice exercises. *Journal of Educational Computing Research*, 5(1), 89-114.

Frick, T. W. (1990). A comparison of three decision models for adapting the length of computer-based mastery tests. *Journal of Educational Computing Research*, 6(4), 479-513.

Frick, T. W. (1992). Computerized adaptive mastery tests as expert systems. *Journal of Educational Computing Research*, 8(2), 187-213.

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

Grégoire, J. (Ed.). (1996). *Evaluer les apprentissages : Les apports de la psychologie cognitive*. Bruxelles : De Boeck.

Guskey, T. R. (1987). The essential elements of mastery learning. *Journal of Classroom Interaction*, 22(2), 19-22.

Guskey, T. R., & Pigott, T. D. (1988). Research on group-based mastery learning programs : A meta-analysis. *Journal of Educational Research*, 81(4), 197-216.

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

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

Holtzman, W. H. (Ed.). (1970). *Computer-assisted instruction, testing, and guidance*. New York : Harper and Row.

Houssemand, C. (2001). *Adaptabilité stratégitaire dans la résolution des Cubes de Kohs*. Lille : Presses Universitaires du Septentrion.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). *Item response theory : Application to psychological measurement*. Homewood : Dow Jones-Irwin.

Ingenkamp, K. (1989). *Diagnostik in der Schule : Beiträge zu Schlüsselfragen der Schülerbeurteilung*. Weinheim : Beltz Verlag.

Kingsbury, G. G., & Houser, R. L. (1999). Developing computerized adaptive tests for school children. In F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in computerized assessment* (pp. 93-115). Mahwah, NJ : Lawrence Erlbaum Associates.

Kingsbury, G. G., & Weiss, D. J. (1983). A comparison of IRT-based adaptive mastery testing and a sequential mastery testing procedure. In D. J. Weiss (Ed.), *New horizons in testing : Latent trait test theory and computerized adaptive testing* (pp. 257-283). New York : Academic Press.

Kulik, C. I. C., Kulik, J. A., & Bangert Drowns, R. L. (1990). Effectiveness of mastery learning programs : A meta-analysis. *Review of Educational Research*, 60(2), 265-299.

Kyllonen, P. C. (1997). Smart testing. In R. F. Dillon (Ed.), *Handbook on testing* (pp. 347-368). Westport, CT : Greenwood Press.

Lewis, C., & Sheehan, K. (1990). Using Bayesian decision theory to design a computerized mastery test. *Applied Psychological Measurement*, 14(4), 367-386.

Liefeld, J. P., & Herrmann, T. F. (1990). Learning consequences for university students using computerized mastery testing. *Educational Technology Research and Development*, 38(2), 19-25.

Lord, F. M. (1970). Some test theory for tailored testing. In W. H. Holtzman (Ed.), *Computer-assisted instruction, testing, and guidance* (pp. 139-183). New York : Harper and Row.

Lord, F. M. (1971a). Tailored testing, an application of stochastic approximation. *Journal of the American Statistical Association*, 66, 707-711.

Lord, F. M. (1971b). A theoretical study of two-stage testing. *Psychometrika*, 36(3), 227-242.

Lord, F. M., & Novick, M. R. (1968). *Statistical theory of mental test scores*. Reading, MA : Addison-Wesley.

Luecht, R. M. (1996). Multidimensional computerized adaptive testing in a certification or licensure context. *Applied Psychological Measurement*, 20(4), 389-404.

Martin, R. (1997). Possibilités d'utilisation du comportement exploratoire dans un espace virtuel en 3D en tant que mesure de l'intelligence spatiale. In J. Juhel & T. Marivain & G. Rouxel (Eds.), *Psychologie et différences individuelles : Questions actuelles* (pp. 69-74). Rennes : Presses Universitaires de Rennes.

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

Martin, R., & Houssemand, C. (2002). Intérêts et limites de la chronométrie mentale dans la mesure psychologique. *Bulletin de Psychologie*, 56.

Psychologie et Psychométrie. Vol. 24 - N° 2/3 - 2003

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests : A meta-analysis. *Psychological Bulletin*, 114(3), 449-458.

Olsen, J. B., Maynes, D. D., Slawson, D., & Ho, K. (1989). Comparisons of paper-administered, computer-administered and computerized adaptive achievement tests. *Journal of Educational Computing Research*, 5(3), 311-326.

Orey, M. A., & Nelson, W. A. (1994). Development principles for intelligent tutoring systems : Integrating cognitive theory into the development of computer-based instruction. *ETR&D - Educational Technology Research & Development*, 41, 59-72.

Owen, R. J. A. (1975). A bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, 70, 351-356.

Reckase, M. D. (1983). A procedure for decision making using tailored testing. In D. J. Weiss (Ed.), *New horizons in testing : Latent trait test theory and computerized adaptive testing* (pp. 237-255). New York : Academic Press.

Richard, F., & Zamani, M. (1996). L'application des modèles de résolution de problème à l'analyse des tests. *Psychologie Française*, 14(1), 77-88.

Rouiller, Y. (1998). L'évaluation formative à l'école : Quelle place pour la régulation interactive ? *Travaux neuchâtelois de linguistique*, 29, 119-133.

Scriven, M. (1967). The methodology of evaluation. *AERA Monograph Series on Evaluation*, 1, 39-83.

Tatsuoka, K. K. (1983). Rule space : An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement*, 20(4), 345-354.

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

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

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

Wainer, H. (2000b). Rescuing computerized testing by breaking Zipf's law. *Journal of Educational and Behavioral Statistics*, 25(2), 203-224.

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

Wald, A. (1947). *Sequential analysis*. New York : Wiley.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, 21(4), 361-375.

Welch, R. E., & Frick, T. W. (1993). Computerized adaptive testing in instructional settings. *Educational Technology Research and Development*, 41(3), 47-62.

Wise, S. L., & Kingsbury, G. G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicologica*, 21(1-2), 135-155.

Xiao, B. (1999). Strategies for computerized adaptive grading testing. *Applied Psychological Measurement*, 23(2), 136-146.

Zara, A. R. (1999). Using computerized adaptive testing to evaluate nurse competence for licensure : Some history and forward look. *Advances in Health Sciences Education*, 4(1), 39-48.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Cambridge, MA : Addison-Wesley.

Psychologie et Psychométrie. Vol. 24 - N° 2/3 - 2003

La méthodologie des évaluations internationales de compétences

Thierry Rocher

Summary : *Numerous international studies have sought to compare the attainments of pupils in various subject areas. Most of these studies rest on a common method which is developed and refined from one study to another. PISA (Programme for International Student Assessment) takes benefits from these improvements. This common method is strongly influenced by the preoccupation of ensuring the construction of a league table that is as fair as possible. Thus the question of the comparability of the results is limited to that of the robustness of the league table. But, in spite of efforts made to make the measuring instruments universally valid, many "cultural" biases remain and weaken the reliability of the final league table. However, far from invalidating the results of international assessments, these biases often bring interesting insights into the attainments compared across pupils that national assessments would not have revealed. Finally, beyond the comparability of the results obtained in terms of ranking, the comparability from the point of view of the dispersion would certainly deserve to be looked at in depth, since the study of inequalities constitutes an important part of international assessments. This subject is a delicate question which relates as much to the adaptability of the test as to the choice of model which may influence the results.*

Key words : *International comparisons, assessments, students' competencies, item response models.*

* Chargé d'études statistiques, Direction de l'Evaluation et de la Prospective du Ministère de l'Education nationale. 3/5 Boulevard Pasteur, 75015 PARIS.