Mesure et évaluation en éducation

Réflexions sur la modélisation statistique des scores aux tests

Jean-Guy Blais and Michel Fournier

Volume 19, Number 3, 1997

Variations culturelles sur le thème ADMEE

URI: https://id.erudit.org/iderudit/1091546ar
DOI: https://doi.org/10.7202/1091546ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

0823-3993 (print)
2368-2000 (digital)

Explore this journal

Cite this article

Blais, J.-G. & Fournier, M. (1997). Réflexions sur la modélisation statistique des scores aux tests. *Mesure et évaluation en éducation*, *19*(3), 69–93. https://doi.org/10.7202/1091546ar

Article abstract

This paper is proposing a reflexion on statistical modeling of tests scores. It is questionning the real differences that one could observe between "Classical test theory" and "Item response theory". It maintains that these two views of testing are useful for classification and prediction of those who take large-scale tests, but little useful for diagnosis of examinees' problems. It ends by suggesting that more fruitfull ways of seeing the problem are getting more and more attention. These promising approaches try to merge test theories and findings in cognitive psychology and focus their attention on process rather than solely on results.

Mesure et évaluation en éducation
Vol. 19 n° 3, pages 69 à 93.

Réflexions sur la modélisation statistique des scores aux tests

Jean-Guy Blais et Michel Fournier
Université de Montréal

Cet article propose une réflexion sur la modélisation statistique des scores aux tests. Il s'interroge sur les différences réelles qui existent entre la « Théorie classique des tests » et la « Théorie de la réponse à l'item ». Il avance que ces deux approches sont utiles pour prédire et classer des candidats qui subissent des tests en grands groupes, mais très peu utiles pour le diagnostic pédagogique. Il termine en suggérant que de nouvelles perspectives s'ouvrent aux praticiens. Ces perspectives s'intéressent principalement aux processus et puisent leur inspiration dans le développement concerté des théories des tests et de la psychologie cognitive.

(théorie classique des tests, théorie de la réponse à l'item, sélection, psychologie cognitive, diagnostic)

This paper is proposing a reflexion on statistical modeling of tests scores. It is questionning the real differences that one could observe between « Classical test theory » and « Item response theory ». It maintains that these two views of testing are useful for classification and prediction of those who take large-scale tests, but little useful for diagnosis of examinees' problems. It ends by suggesting that more fruitfull ways of seeing the problem are getting more and more attention. These promising approaches try to merge test theories and findings in cognitive psychology and focus their attention on process rather than solely on results.

(classical test theory, item response theory, selection, cognitive psychology, diagnosis)

Introduction

La modélisation des scores aux tests en éducation et en psychologie privilégie depuis le début du siècle un paradigme statistique. On s'est peu intéressé, sinon d'une façon parcimonieuse, à une compréhension de l'interaction entre l'item et le candidat qui permettrait un meilleur diagnostic des mécanismes de production des performances observées.

Dans la modélisation classique, les items d'un test n'ont pas vraiment de caractère propre : ils font partie du test et leurs propriétés métrologiques en sont dépendantes. On les étudie individuellement uniquement pour s'assurer qu'ils visent tous la même direction (tout en recherchant une discrimination élevée), mais ils ne peuvent pas être considérés isolément des autres items avec lesquels ils ont été calibrés. Cette modélisation est avant tout statistique et sert à déterminer certaines caractéristiques métrologiques du test, comme son degré de fidélité ou de consistance interne. Si cette approche peut se montrer efficace lorsqu'elle est utilisée dans des situations de sélection et de placement, elle est peu efficace lorsqu'il s'agit de diagnostiquer les causes des problèmes individuels d'apprentissage.

Depuis une vingtaine d'années, une modélisation opposée, axée non plus sur le test mais sur l'item, a permis l'émergence d'un nouveau complexe de *testing*. Cette nouvelle modélisation porte le nom maintenant consacré de « Théorie de la réponse à l'item » (*Item Response Theory*).

La modélisation de la théorie de la réponse à l'item isole l'item et le candidat des entités auxquelles ils sont reliés dans la théorie classique des tests, c'est-à-dire le test et l'échantillon de candidats. C'est ce qui distingue nettement la nouvelle modélisation : elle s'intéresse au résultat généré par la rencontre entre un item donné et un candidat donné. L'item peut, en principe, avoir une existence propre, isolée de celle des autres items avec lesquels il est utilisé. C'est également le fait d'utiliser un modèle de régression qui amène un élément de nouveauté et qui engendre une propriété théorique importante : la propriété d'invariance, qui se manifeste de deux façons différentes. D'abord, les valeurs des paramètres d'un item sont indépendantes de l'ensemble des items avec lesquels il est calibré. Ensuite, l'estimation de l'habileté d'un candidat est indépendante de l'échantillon de candidats servant à obtenir l'estimation.

Toutefois, la nouvelle modélisation n'est pas incompatible avec l'ancienne et les paramètres des modèles proposés symbolisent des cactéristiques semblables à celles qui existaient dans la théorie classique des tests. On peut en effet retrouver, dans les modèles les plus utilisés actuellement avec les variables cognitives (les modèles de Rasch et les modèles logistiques avec deux ou trois paramètres), des paramètres associés à la difficulté et au pouvoir de discrimination d'un item et un paramètre associé à l'habileté du candidat.

La nouvelle modélisation est perçue par certains comme une révolution dans le domaine de la mesure en éducation. Ainsi, Wright (1988) et Reckase

(1989) présentent cette modélisation en des termes qui ne laissent planer aucun doute quant à la place qu'ils estiment qu'elle occupe dans le développement de la mesure :

> Les modélisations proposées par Rasch constituent les progrès les plus importants à survenir en psychométrie depuis la *Loi des jugements comparés* de Thurstone en 1927. (Wright, 1988, p. 286.)

> La mesure a atteint un nouvel âge d'or. Le nombre de problèmes intéressants et d'approches pour les résoudre sont presque sans limites. Un élément central des développements récents dans la technologie de la mesure est l'accent mis sur l'interaction entre la personne et l'item. C'est là une retombée du développement de la méthodologie de la théorie de la réponse à l'item. L'attention portée à cette interaction par la communauté des spécialistes de la mesure a produit des développements analogues aux développements en chimie et en physique durant le 19e siècle à la suite du développement de la théorie atomique. (Reckase, 1989, pp. 14-15.)

Mais à quelle sorte de progrès ou de supériorité fait-on allusion? Est-ce que les changements permettent d'assimiler la nouvelle modélisation à une révolution au sens de l'épithète consacrée par Thomas Kuhn dans *La structure des révolutions scientifiques* (1983)? Avons-nous vraiment assisté à un changement de paradigme dans le domaine de la mesure en éducation et en psychologie ou bien est-ce la poursuite de la science que Kuhn qualifie de « normale », c'est-à-dire, dans le cas qui nous intéresse, la continuité de la domination d'une conception empirique de la mesure?

Le présent article veut montrer que les modélisations classique et nouvelle appartiennent à une même vision de la mesure. Cette vision est plutôt empirique et la modélisation conceptuelle y joue un rôle secondaire. Ces modélisations poursuivent des objectifs de mesure puisant leurs origines dans un virage pris par la faction quantitative de la psychologie différentielle au 19e siècle. Elles ont démontré leur efficacité pour ordonner les élèves et remplir des fonctions de sélection et de prédiction, mais elles ont également démontré très peu d'utilité pour fournir un diagnostic ou pour justifier un changement de stratégie pédagogique (la nouvelle modélisation est plus flexible à ce titre que l'ancienne). Également, cet article décrit brièvement certains travaux récents fusionnant le *testing* et la psychologie cognitive. Ces travaux représentent une façon différente d'envisager le *testing* et ils pourraient paver la voie à un apport conceptuel basé sur une réflexion théorique au service du diagnostic et de la compréhension.

L'influence de la psychologie différentielle : prédire et sélectionner

Au 17e siècle, le triomphe de la physique galiléenne quantitative sur la physique aristotélicienne qualitative a pavé la voie à suivre pour les développements de la science pratiquement jusqu'à nos jours. Les succès de Newton et de ses successeurs dans l'explication de phénomènes physiques au moyen de théories où la quantification jouait un rôle prédominant, que ce soit la gravitation, la chaleur, l'électricité ou le magnétisme, confortèrent la thèse selon laquelle la science était synonyme de science quantitative. L'entreprise scientifique au cours des 18e et 19e siècles peut être décrite comme une extension de la quantification à de nouveaux domaines, conjuguée à une recherche de la précision (Frängsmyr et al.,1990; Wise, 1995).

Dans cet esprit, une psychologie scientifique naissante dans la seconde moitié du 19e siècle, qui tentait de se démarquer de la psychologie philosophique, devait donc se plier à cet impératif quantitatif (Michell, 1990). Or, la production de données chiffrées dépendait de la possibilité de créer des instruments de mesure appropriés. Mais avec quel mètre, sur quelle balance allait-on pouvoir jauger les différents et intangibles aspects de l'esprit humain? Ainsi, c'est en filiation avec des préoccupations concernant la mesure des différences individuelles que s'est développé ce qu'on a distingué comme les théories des tests d'habiletés intellectuelles (*mental test theories*, Boring, 1961).

Quételet avait montré, en 1835, que la distribution de la taille des soldats de l'armée napoléonienne suivait approximativement la courbe d'erreur normale utilisée par les astronomes. Galton postula que l'intelligence était répartie de la même façon et tenta de l'estimer, d'abord au moyen de la réputation des individus, puis par des tests composés en majeure partie d'épreuves sensorielles (acuité visuelle et auditive, discrimination des couleurs, évaluation de distances, temps de réaction). En vertu de la théorie associationniste qui soutenait que les idées sont le résultat de combinaisons de sensations élémentaires, les sujets possédant les sens les plus fins devaient également être les plus brillants (Kendler, 1987).

Ces tests, introduits aux États-Unis par J.M. Cattell (1860-1944) dans la dernière décennie du 19e siècle, donnèrent lieu à des études sur une échelle importante. On s'aperçut cependant rapidement que les résultats à ces épreuves sont faiblement liés à la réussite scolaire ou universitaire (Reuchlin, 1989).

Au même moment, de l'autre côté de l'Atlantique, Alfred Binet (1857-1911) réalisait une série de tests visant à évaluer directement les facultés psychiques supérieures (mémoire, jugement, raisonnement, compréhension, etc.). Constitués d'un amas d'épreuves hétéroclites, ne faisant pas appel aux techniques apprises, comme la lecture, ces tests étaient rangés dans un ordre croissant de difficulté, auquel était associé l'âge où un enfant d'intelligence normale devait pouvoir les réussir. L'objectif de Binet était de distinguer les enfants anormaux de façon à ce que ceux-ci puissent bénéficier d'un encadrement qui leur conviendrait mieux que la filière régulière (Schneider, 1992). Il favorisait le diagnostic des problèmes et la mise en place de solutions individuelles.

Importée aussitôt aux États-Unis, la démarche de Binet fournit des pronostics qui se révélèrent bien supérieurs à ceux des tests de Galton-Cattell. Ce fut le début d'un nouvel essor de la psychologie différentielle, mais aussi de son détournement plus ou moins conscient à des fins idéologiques. Car, si Binet avait conçu son échelle essentiellement dans un but diagnostique, elle fut néanmoins traitée par H.H. Goddard (1866-1957) comme une évaluation directe de l'intelligence innée. L.M. Terman (1877-1956) en élabora bientôt une nouvelle version, l'échelle Stanford-Binet, et se fit l'avocat de son utilisation universelle dans le but d'écarter de la société ceux dont l'intelligence est trop faible, mais surtout de les empêcher de se reproduire (point de vue soutenu par les défenseurs de l'eugénisme). Clairement, l'objectif visé était de sélectionner les meilleurs et de prédire la réussite future.

Lors de la première guerre mondiale, R.M. Yerkes (1876-1956), avec une méthode que Gould (1983) qualifie de douteuse, fit subir ses tests *Alpha* et *Bêta* aux quelque 1 700 000 recrues de l'armée américaine. Les conclusions de l'analyse des données furent publiées en 1921 et menèrent directement au *Johnson-Lodge Immigration Act* de 1924 (Thomson & Sharp, 1988), qui restreignait sévèrement l'entrée aux États-Unis d'immigrants en provenance de contrées considérées comme intellectuellement défavorisées. Le Q.I. — mesure dérivée des tests de Binet — devenait un argument à des thèses héréditaristes et eugénistes, une forme de justification rationnelle au racisme.

Justification qui apparaissait d'autant plus inattaquable qu'elle pouvait s'appuyer sur de nouvelles techniques statistiques objectives. Afin de vérifier le caractère héréditaire de ses mesures, Galton avait été amené à créer le concept de la corrélation entre deux variables. K. Pearson (1857-1936), lui donnant une expression mathématique convenable, en fera le coefficient de corrélation *r*. Par la suite, ce même concept, étendu à plusieurs dimensions

en 1903 par C.E. Spearman (1863-1945) sous le nom d'analyse factorielle, servira à trouver un facteur commun sous-jacent aux diverses épreuves des tests de Binet, le facteur *g* ou intelligence générale. À ce sujet, Gould (1983) a écrit :

> Avec ces principes, écrivit Spearman en 1923, il faut se hasarder à espérer que l'on ait pu pourvoir la psychologie de ce fondement qui lui faisait défaut depuis si longtemps, de manière à ce que désormais elle prenne sa place légitime parmi les autres sciences solidement établies, comme la physique elle-même (p. 328).

De l'analyse factorielle elle-même allaient cependant surgir des objections quant à la validité du facteur *g*. Dans *Vectors of the mind* (1935), L.L. Thurstone (1887-1955) proposa une variante à la méthode de Spearman, qui fait passer le nombre d'aptitudes mentales primaires à sept. Guilford, en 1936, reconnaît, pour sa part, cent vingt facteurs différents. Selon ces approches concurrentes, la notion d'intelligence générale cède la place à une multitude d'aptitudes distinctes.

Ainsi, par divers sentiers, la nouvelle psychologie de la fin du 19e siècle s'efforçait tant bien que mal de quantifier le domaine intangible des facultés humaines, se détachant ainsi graduellement de la philosophie de l'âme, où l'on prônait l'introspection et la spéculation subjective, pour tenter d'accéder au rang de science.

Les débuts du *testing* en milieu scolaire

À la fin du 19e siècle, la psychologie, forte de son tout récent statut scientifique, va s'institutionnaliser. Elle formera ses diplômés, créera ses propres revues, ses laboratoires et, dans un même mouvement, elle cherchera à s'introduire un peu partout, notamment à l'école. N'avait-on pas besoin là, comme ailleurs, des lumières de la science? Ne pouvait-on pas y implanter des procédures fondées rationnellement? Danziger (1987, p. 40) décrit d'ailleurs la relation à cette époque entre la psychologie et l'éducation comme ayant les caractéristiques d'une symbiose.

Les méthodes de la mesure des différences individuelles font d'abord leur entrée dans le système éducatif, avec Binet en France, pour le dépistage des « faibles d'esprit ». Puis, à partir des années 20 aux États-Unis, l'usage des tests standardisés se répand pour la **sélection** et le **classement** des élèves dits normaux (Thomson & Sharp, 1988). Il faut souligner qu'à cette époque, les

États-Unis vivaient de profonds changements qui étaient la conséquence directe de la révolution industrielle de la fin du 19e siècle. La poursuite d'un objectif de scolarisation du plus grand nombre possible d'enfants (*mass education system*) amenait des contingences à l'admission dans les écoles. D'un système d'éducation axé sur les privilèges d'une poignée d'élus, on passa ainsi à un système basé sur le mérite, une sorte de méritocratie (Evans & Waites, 1981). Il fallait donc trouver une stratégie efficace pour gérer la masse grandissante d'élèves frappant aux portes des écoles et déterminer ceux qui méritaient d'être admis. Les tests d'habiletés intellectuelles, plus précisément les tests d'intelligence, trouvèrent donc leur niche parce que les impératifs sociaux de l'époque exigeaient des outils pouvant contribuer à effectuer efficacement un travail de classement et de sélection.

De la psychologie différentielle naît donc une sous-discipline, la **théorie des tests**, vouée à l'application de ses principes en milieu scolaire. Les revues créées pour la diffusion des résultats de la recherche portent à cet égard des titres significatifs (tableau 1). Si d'emblée ces revues publient indifféremment des articles psychométriques de diverses natures, il faut attendre les années 60 avant de voir un périodique traiter exclusivement de la mesure en éducation, témoignage patent de la croissance et de la différenciation progressive de la sous-discipline.

Tableau 1

Première parution de journaux consacrés à la mesure en psychologie et en éducation

Revues	Première parution
Psychometrika	1936
Educational & Psychological Measurement	1941
British Journal of Mathematical & Statistical Psychology	1948
Journal of Educational Measurement	**1964**
British Educational Research Journal	1975
Journal of Educational Statistics	1976
Applied Psychological Measurement	1977
Mesure et évaluation en éducation	1978
Educational Measurement : Issues & Practices	1982
Applied Measurement in Education	1988

En dehors de toute considération physiologique, la théorie des tests, à l'instar de la psychologie différentielle, cherchera à inférer, à partir de réponses à des questions (les items des tests), des résultats qui se veulent représentatifs d'une certaine habileté des individus, habileté dont la nature comme l'action demeurent indéterminées. Le but n'est pas de comprendre l'intellect, mais de **prédire** ses performances. Quant à la validité de ces mesures, l'attitude adoptée par les chercheurs paraîtra souvent ambiguë. Elle balance entre une forme de pragmatisme pour lequel une théorie n'est bonne que dans la mesure où elle se montre utile, et une tendance à la réification où les lois statistiques deviennent, en quelque sorte, autonomes (Hacking, 1987, p. 53). L'étude de certains concepts comme la multidimensionalité d'un test ou encore le fonctionnement différentiel d'un item, permet d'ailleurs de se rendre compte que cette dernière tendance est toujours présente chez certains psychométriciens de notre époque.

Issue directement de la psychologie différentielle, la théorie des tests sera élaborée souvent par les mêmes chercheurs — Spearman, par exemple — qui publieront leurs travaux dans les mêmes revues — *Psychometrika* a été fondée par Thurstone — et porteront un intérêt identique aux méthodes corrélatives. Oeuvre de statisticiens, plus que de psychologues ou d'éducateurs, isolée des autres grands courants de la psychologie, la mesure des variables cognitives en éducation apparaît comme une cathédrale statistique élevée sur des postulats psychologiques minimalistes.

Comme le dit si bien Mislevy (1993) :

> Ce n'est qu'une légère exagération que de décrire la théorie des tests qui domine aujourd'hui la mesure en éducation comme l'application de la statistique du 20e siècle à la psychologie du 19e siècle (p. 19).

La théorie classique des tests

L'information fournie par les scores aux tests est avant tout de nature ordinale. Les items ne sont pas toujours équivalents d'un point de vue métrique et l'addition des scores aux items ne donne pas automatiquement naissance à une échelle où les distances entre les scores sont équivalentes. La grande ambition des modélisations utilisées en éducation sera donc de transformer cet ordre en une échelle à intervalles, où l'on pourra apprécier la distance entre différents résultats.

Créateur de l'analyse factorielle, Spearman élabora également, dans des écrits publiés entre 1904 et 1913, les bases de la première tentative en ce sens. L'idée centrale de la *théorie classique des tests* est que la corrélation entre deux tests visant à évaluer un même attribut psychique (qui sont dits tests parallèles) peut servir à contrer les effets des erreurs de mesure, supposées aléatoires, et fournir une estimation de la fidélité de la mesure que constituent les scores à un test.

La théorie classique des tests fut élaborée, reformulée et axiomatisée, pour reprendre les termes de Lumsden (1976), par une lignée royale d'auteurs (Thurstone, 1931; Guilford, 1936; Gullicksen[1], 1950; Magnusson, 1967; Lord & Novick, 1968). Mais sous quelque appellation qu'on lui donne (coefficients de stabilité, d'équivalence, de précision ou de consistance interne) c'est toujours le concept de corrélation que l'on retrouve et la même intention d'obtenir une estimation de la fidélité de la mesure à partir d'une modélisation statistique[2].

Théorie peu restrictive dans ses postulats, la théorie classique des tests n'en présente pas moins un certain nombre de défauts qui vont éventuellement pousser les psychométriciens à la recherche de nouvelles modélisations. Par exemple, Hambleton et Swaminathan (1985) soulignent les inconvénients suivants :

- la fidélité d'un test est directement reliée à la variance des scores;
- la notion cruciale de tests parallèles est difficile à réaliser en pratique;
- elle présume que la variance des erreurs de mesure est la même pour tous les candidats.

De plus, l'application de la théorie classique des tests aux résultats d'un test peut mener à des paradoxes (Cliff, 1989). Par exemple, le paradoxe de l'atténuation, où l'accroissement de la consistance interne d'un test cause une diminution de sa corrélation avec ce qu'on appelle le score-vrai (*true score*). De sorte qu'au cours des années 60, le sentiment d'insatisfaction envers la théorie classique des tests ira croissant (Bock & Wood, 1971, pp. 197-198).

Les psychométriciens porteront de plus en plus leur attention vers des modélisations pouvant suppléer à la théorie classique des tests : la théorie de la généralisabilité (Cronbach, Gleser & Rajaratnam, 1963), les théories fortes du score-vrai (Lord, 1965), les modèles d'échantillonnage des items (Lord &

Novick, 1968) et surtout une modélisation dont la complexité mathématique avait jusqu'alors entravé les applications : la théorie de la réponse à l'item.

Il convient peut-être de souligner plus amplement les efforts faits par les promoteurs de la théorie de la généralisabilité pour libéraliser la théorie classique des tests. La théorie de la généralisabilité a mis à profit les développements importants de l'analyse de la variance en statistique pour proposer une vision différente de l'erreur de mesure. Plutôt que de considérer l'erreur de mesure comme une entité globale, la théorie de la généralisabilité conçoit les différentes sources d'erreur comme des entités séparées, dont on peut étudier les influences sur la variation des scores observée. La théorie de la généralisabilité met l'accent sur l'estimation des différentes composantes de la variance des scores et ne postule pas l'existence de traits latents non observables (comme la théorie de la réponse à l'item). Ses objectifs sont cependant restés relativement les mêmes que ceux de la théorie classique des tests[3].

La théorie de la réponse à l'item

La théorie de la réponse à l'item présume l'existence dans l'esprit humain de variables psychologiques non observables et relativement stables (les traits latents ou encore les habiletés) pouvant expliquer les scores observés à un test. Elle suppose de plus que les réponses aux différents items sont statistiquement indépendantes : c'est le postulat d'indépendance locale. Afin d'obtenir une mesure de l'un de ces traits — on parlera alors d'un modèle unidimensionnel — on exprime la probabilité de réussite, pour chacun des items, en fonction du niveau d'habileté. Pour ce faire, on choisit une famille de courbes, bornées entre 0 et 1 sur l'ordonnée (puisqu'il s'agit d'une probabilité de succès), et on estime la valeur des paramètres de sorte que la courbe décrive le mieux possible la réponse à un item particulier. Cette courbe caractéristique de l'item représente la régression du score sur la variable latente et, de ce fait, lorsque la situation se conforme aux exigences du modèle, les résultats obtenus sont théoriquement invariants (Blais & Ajar, 1992).

Les origines de cette modélisation peuvent être retracées au tout début de la théorie des tests. En effet, Binet et Simon avaient, dès 1916, présenté des tableaux de la proportion de réponses correctes à un item en fonction de l'âge. De même, Thurstone, en 1925, avait rangé des items selon une échelle d'âge correspondant à cinquante pour cent de succès (Goldstein & Wood, 1989).

C'est Ferguson (1942) qui introduit ce qui inspirera la façon actuelle de présenter la relation entre l'item et l'individu. Pour Ferguson, la courbe obtenue représente la *probabilité* qu'une personne d'une habileté donnée réponde correctement à un item particulier, et les paramètres de l'item sont reliés à la difficulté et à la discrimination de l'item. Cette formulation probabiliste sera présente dans tous les développements ultérieurs de la théorie de la réponse à l'item, mais les chercheurs donneront du processus aléatoire en jeu des interprétations variées. Par exemple, selon Holland (1990), cette probabilité pourrait être définie d'au moins trois façons :

1. L'échantillonnage des sujets : la proportion des individus d'une habileté A qui réussissent un item I.

2. Un sujet stochastique : la probabilité qu'un individu X réponde correctement à un item I.

3. L'échantillonnage des items : la proportion des items, ayant la même courbe caractéristique, qu'un individu X pourrait réussir.

Lawley (1943) appliqua au problème de l'estimation des paramètres un traitement différent en employant la méthode de vraisemblance maximale de Fisher. Les équations auxquelles il parvint n'étant pas résolubles analytiquement, il dut restreindre son étude à des items équivalents et recourir à des approximations numériques. Il montra aussi comment les paramètres obtenus pouvaient être reliés aux concepts de la théorie classique des tests, travail poursuivi par la suite par Tucker (1946) pour le cas particulier où l'habileté est distribuée normalement dans la population. Le nouveau paradigme, comme on le constate, ne s'est pas encore affranchi de ses liens avec la théorie dominante du moment. Il se réfère toujours à la théorie classique des tests pour se justifier, pour donner un sens à ses résultats.

Avec Lord (1952, 1953), qui établit la jonction avec les classes latentes de Lazarsfeld (1950) utilisées en sociologie, la modélisation prend le nom de *Théorie du trait latent*. Rassemblant les travaux antérieurs, et précisant leurs fondements statistiques, notamment le postulat d'indépendance locale, il fut en mesure de bâtir une théorie-modèle consistante. Ce modèle comprenait en outre des solutions à différents problèmes de *testing* : pondération des items, puissance discriminatoire pour différents niveaux d'habileté et pour différents scores au test, intervalles de confiance autour du score-vrai et distribution optimale de la difficulté des items. À ce point arrivèrent du domaine de la toxicologie des développements statistiques qui devaient faciliter la tâche des psychométriciens. Berkson et Anscombe avaient en effet

montré, au début des années 50, que la loi de probabilité logistique représentait une solution de rechange avantageuse à la loi de probabilité normale (Maxwell, 1959).

La publication de *Statistical theories of mental test scores* (Lord & Novick, 1968) devait assurer le triomphe de la fonction logistique. Birnbaum y propose un modèle qui, comme celui de Lord, exige deux paramètres, la difficulté et la discrimination, pour décrire la réponse à un item. Un troisième paramètre, la pseudo-chance, est introduit pour rendre compte des situations où les sujets peuvent deviner la bonne réponse.

Le modèle le plus simple, un modèle logistique ne recourant qu'à un seul paramètre, soit la difficulté, fut proposé de manière tout à fait indépendante par Georg Rasch en 1960. Le modèle de Rasch rend plus exigeante la sélection des items du test, car il demande que leurs discriminations soient égales[4] et que la pseudo-chance soit un facteur négligeable. Conséquemment, le modèle de Rasch est utilisé comme un modèle normatif plutôt que descriptif. Ainsi, les items sont choisis en fonction de leur adéquation au modèle, et non pas l'inverse. Il en résulte à la fois un accroissement de l'homogénéité du test et une perte possible de la représentativité du contenu (Wood, 1978).

Une autre propriété, liée à la précédente, distingue le modèle de Rasch de celui de Birnbaum : l'habileté du sujet et la difficulté de l'item peuvent y être exprimées sur une même échelle. C'est ce que les propagateurs du modèle de Rasch appelle l'objectivité spécifique. La pertinence de cette propriété pour la mesure des traits latents va diviser la communauté psychométrique en deux camps opposés. D'un côté, ceux qui la jugent superflue, ne verront dans le modèle de Rasch qu'un cas particulier du modèle logistique à trois paramètres, qu'une modélisation certes plus simple, mais dont les postulats restent plus contraignants (Divgi, 1989, pp.298-299; Whitely, 1977, p. 233). D'un autre côté, certains allèguent que l'objectivité spécifique représente une caractéristique essentielle de la mesure. D'abord, parce qu'elle obéit au principe de la parcimonie scientifique, qui veut qu'on explique un phénomène à l'aide du plus petit nombre de variables possible; ensuite, parce que, répondant aux conditions de la mesure conjointe additive telle que définie par Luce et Tuckey (1964) et Fischer (1987), elle procure une mesure fondamentale, c'est-à-dire qu'elle implique une structure additive similaire à la concaténation de tiges bout à bout (Wright, 1984, p. 283).

Les dissensions évoquées en ce qui a trait aux qualités désirables d'une mesure révèlent l'existence, au sein de la communauté des chercheurs en

psychométrie, de deux visions de la théorie de la réponse à l'item qui, à première vue, semblent antagonistes. Toutefois, la littérature polémique est peu abondante, les chercheurs s'appliquent surtout à développer la théorie sous un angle statistique et les applications visent le plus souvent les mêmes objectifs : sélectionner, classer, prédire.

La théorie de la réponse à l'item : une révolution ou une poursuite de la science normale?

En 1976, Lumsden réclamait l'abandon pur et simple de la théorie classique des tests, qualifiant au passage la tentative de rassemblement et d'axiomatisation de ses résultats accomplie par Lord et Novick (1968) de grandiose mausolée. Pourtant, constatent Weiss et Davison (1981), au cours de la période 1975-1979, alors que de nouvelles modélisations sont disponibles, la recherche sur la théorie classique se poursuit sans fléchir. Toutefois, dans leur revue de la documentation parue entre 1980 et 1982, Traub et Lam (1985) ne traitent même plus de la théorie classique des tests. Il semble bien que sa popularité a décliné chez les psychométriciens et que des modèles plus récents, dont au premier chef la théorie de la réponse à l'item, l'ont reléguée à l'arrière-plan.

Si, au cours des années 50 et 60, la complexité mathématique de la théorie de la réponse à l'item avait rendu ses progrès d'une lenteur pénible (Hambleton & Swaminathan, 1985, p.7), à partir de 1968, la cadence s'accélère, le nombre d'articles publiés annuellement sur le sujet s'accroît de manière spectaculaire. Alors qu'on pouvait compter une poignée d'articles en 1970, on en retrouve plus d'une quarantaine vingt ans plus tard (Fournier, 1994). Outre les carences déjà signalées de la théorie classique des tests, on peut s'interroger sur les causes de la faveur conjointe qu'obtiennent alors les modélisations de Rasch et de Birnbaum — les paradigmes siamois, pourrait-on dire.

En premier lieu, l'avènement des ordinateurs a rendu praticables les fastidieux calculs d'estimation des paramètres. Dès 1959, Baker décrivait un programme informatique en ce sens et entrevoyait l'impact qu'allaient avoir ces machines sur la théorie des tests (Baker, 1959, p. 242). Ensuite, les modèles de Birnbaum et de Rasch, avec les procédures d'estimation itératives qui leur sont associées, ont pu jouir d'une large diffusion : le premier par la publication de Lord et Novick (1968), le second grâce au prosélytisme de Wright (1977).

Une édition complète du *Journal of Educational Measurement* consacrée à la théorie de la réponse à l'item en 1977, exemple qui sera répété (tableau 2), va contribuer à la faire connaître davantage encore et marquera un point d'inflexion important dans la production annuelle d'articles sur le sujet.

Tableau 2

Éditions spéciales de périodiques portant sur la théorie de la réponse à l'item

Année	Thème du numéro spécial	Revue
1977	Théorie de la réponse à l'item	Journal of Educational Measurement
1982	Théorie de la réponse à l'item	Applied Psychological Measurement
1984	Ordinateurs personnels et mesure en éducation	Educational Measurement : Issues and Practices
1984	Banque d'items	Journal of Educational Measurement
1986	Banque d'items	Applied Psychological Measurement
1987	Appariement	Applied Psychological Measurement
1989	Les applications de l'ordinateur au *testing*	Educational Measurement : Issues and Practices
1990	Appariement	Applied Measurement in Education

La raison principale qui explique le succès de la théorie de la réponse à l'item reste peut-être la promesse d'invariance qu'elle fait miroiter et l'aboutissement, pour les psychométriciens, de leur longue quête d'une échelle à intervalles. Wright, entre autres, a beaucoup mis l'accent sur ces qualités, en créant, en 1968, les expressions « calibration d'items indépen-

dante des personnes » (*person-free item calibration*) et « mesure des personnes indépendante des items » (*item-free person measurement*). Il est difficile de déterminer si cette invariance alléguée est effectivement atteinte par la théorie de la réponse à l'item au moment où elle effectue sa percée. Même si quelques études tendent à la confirmer (Anderson, Kearney & Everett, 1968; Tinsley & Dawis, 1975, 1977; Slinde et Linn, 1979), l'incertitude quant aux propriétés statistiques des estimateurs, comme le peu d'attention accordée à la vérification des postulats demandent probablement d'éviter les conclusions prématurées concernant la propriété d'invariance. Elle demeure une propriété qui doit être vérifiée empiriquement.

À première vue, ce passage de la théorie classique des tests à la théorie de la réponse à l'item présente plusieurs analogies avec ce que Thomas Kuhn (1983) appelle une révolution scientifique. Ainsi, pourrions-nous définir une période de constitution du nouveau paradigme s'étendant des travaux de Binet jusqu'à la fin des années 60; une phase de transition, où la théorie de la réponse à l'item gagne de plus en plus d'adeptes, qui couvrirait la décennie 70; enfin, l'ère de la théorie de la réponse à l'item triomphante (Hambleton, 1986, p. 415). La phase actuelle, marquée par une extension des modèles, la multiplication des applications et la publication d'ouvrages d'initiation, s'apparente à la phase de science normale qui succède, selon Kuhn, à une révolution scientifique. L'ancien paradigme est détrôné, le nouveau occupe tout l'espace, la révolution est complète.

Mais de quelle sorte de révolution peut-on parler ici? Ne sommes-nous pas plutôt dans la continuité épistémologique de ce qui existait auparavant? Une modélisation empirique, un peu plus sophistiquée du point de vue statistique, mais une modélisation qui porte toujours sur les codes numériques (les plus souvent 0 et 1) associés aux réponses observées. Ne s'agirait-il pas d'une révolution plutôt comparable au remplacement de la règle à calculer par la calculatrice, un changement d'outil plus qu'un bouleversement profond de la manière d'envisager la discipline?

Si la théorie de la réponse à l'item propose certains progrès, certaines améliorations pratiques, elle n'implique pas de transformation dans la vision du monde. Du moins pas sous la forme des modèles les plus simples et les plus populaires actuellement. Elle n'entraîne pas non plus la création de nouvelles institutions ou de nouvelles revues. Loin d'être incommensurables, la théorie classique des tests et la théorie de la réponse à l'item s'expriment dans un même langage et peuvent être ramenées dans un cadre commun (comme le démontrent Goldstein et Wood, 1989). En fait, la théorie de la réponse à l'item ne contredit en rien les postulats de la théorie classique des tests. Elle

rend seulement plus rigoureuse la définition des notions d'unidimensionnalité et d'indépendance des items. La révolution n'est donc, ici, qu'un changement de théorie-modèle.

Un indice supplémentaire que la théorie de la réponse à l'item se situe toujours à l'intérieur de la matrice disciplinaire de la psychologie différentielle nous est donné par les liens récemment établis entre la théorie de la réponse à l'item et l'analyse factorielle et qui nous sont suggérés par McDonald (1986) :

> Je crois que certains psychométriciens qui travaillent sur la théorie de la réponse à l'item ne saisissent pas que ces modèles sont en fait des modèles d'analyse factorielle avec un facteur commun pour les items et constituent une spécialisation d'un modèle non linéaire plus général (p. 518).

Cependant, on trouve peu de témoignages de fidèles qui défendent avec acharnement la bonne vieille théorie classique. Le pragmatisme des psychométriciens a fait en sorte que lorsque la théorie de la réponse à l'item est apparue plus efficace ou, à tout le moins, plus prometteuse que la théorie classique des tests pour résoudre certains problèmes, ils se sont tournés vers celle-ci sans remords, sans avoir l'impression de trahir leur foi.

Paradoxalement, un examen rapide de quelques revues scientifiques où sont publiés les résultats de recherche en éducation et en psychologie montrerait bien que ce que certains considèrent comme une révolution n'a pas atteint la majorité des chercheurs. Ceux-ci font toujours appel aux éléments de la théorie classique des tests, comme le coefficient alpha de Cronbach, pour déterminer les qualités métrologiques de leurs instruments. La révolution s'est peut-être passée dans la tête des psychométriciens, mais elle n'a pas encore atteint les praticiens de la recherche en éducation.

Une véritable révolution?

Ainsi le cycle serait complété. De la crise de la théorie classique des tests, ou à tout le moins de la perte de confiance en ses possibilités, en passant par la longue émergence de la théorie de la réponse à l'item qui se concrétise, dans les années 70, avec des allures de révolution, jusqu'à la phase actuelle que l'on pourrait qualifier de science normale, le développement de la théorie de la réponse à l'item, dans toutes ses étapes, présente des analogies avec les révolutions scientifiques au sens de Kuhn. Force est d'admettre cependant

que la théorie des tests n'a pas été le théâtre d'une authentique révolution kuhnienne.

D'ailleurs la communauté psychométrique, loin d'être monolithique, est divisée sur les mérites du nouveau paradigme proposé. La supériorité de la théorie de la réponse à l'item sur la théorie classique des tests n'est pas reconnue par tous, comme l'affirment Douglass, Khavari et Farber (1979) :

> Si les différences importantes continuent d'être difficilement saisissables, l'efficacité en termes de coût doit être considérée comme la principale distinction entre les deux procédures. Ainsi, le modèle classique apparaît sans compétition comme le modèle de choix lorsqu'une analyse d'items doit être réalisée (p. 351).

Aux affirmations enthousiastes viennent répondre une minorité de critiques tenaces :

> [...] le principe de l'invariance pour des sous-populations [...] est une tautologie qui régit tous les modèles avec un facteur commun. [...] En particulier, le principe d'invariance ne fournit aucune assise pour affirmer que le modèle de Rasch procure des estimations des paramètres des items indépendantes de l'échantillon (McDonald, 1986, p. 518).

En poursuivant l'analogie avec les écrits de Kuhn et en faisant un peu de prospective, on pourrait se demander si la théorie de la réponse à l'item ne se dirige pas à son tour vers une crise. L'indubitable prolifération des modèles proposés au cours des dernières années — on a pu ainsi voir apparaître des modèles dynamiques (Ackerman & Spray, 1986), non paramétriques (Meijer, Sijtma & Smid, 1990), les *testlets* (Wainer & Lewis, 1990), le modèle du quotient (Ramsay, 1989), etc. — pourrait être un signe en ce sens. De l'avis de Golstein et Wood (1989) :

> On peut douter que la prolifération évidente soit un signe de santé. La vraie diversité aurait été bénéfique, mais on a plutôt vu de l'embellissement et du bricolage. Il y a eu très peu d'invention (p. 139).

Cependant, le véritable « challenger » existe. Il a le potentiel pour amener un changement de perspective en proposant un paradigme qui ne s'appuirait pas sur un édifice statistique sophistiqué, mais plutôt sur des développements associés à la psychologie cognitive. Par exemple, certaines tentatives récentes de fusion de la théorie de la réponse à l'item et de théories cognitives (Mislevy, 1993; Embretson, 1983, 1994), donnant une validité de construit aux traits latents et des assises psychologiques à la mesure, pourraient peut-être rompre avec la matrice disciplinaire galtonienne et

constituer une véritable révolution. Pour la première fois depuis le début du siècle, l'élaboration de théories-modèles ne se fait pas en adaptant les réponses des sujets à des modèles statistiques, mais en essayant de comprendre les cheminements qui mènent à la production des réponses. L'objectif n'est plus de mettre en rang et de prédire statistiquement, mais plutôt de comprendre et d'établir un diagnostic à partir de la tâche à réaliser (l'item). Comme le soulignait déjà Whitely (1980) :

> Les modèles de traits latents appliqués sans théorie cognitive, mènent à des tests qui ne mesurent pas des construits cognitifs de base. De la même façon des théories de la performance cognitive qui ne considèrent pas les différences individuelles peuvent n'avoir aucune utilité dans la pratique. Une interaction continue entre la théorie et la modélisation psychométrique mène à des recherches sur l'intelligence qui ont des retombées importantes à la fois au niveau pratique et au niveau théorique (p. 129).

Le contenu de trois monographies récentes permet d'entrevoir plusieurs facettes de ce que pourrait être une évaluation des apprentissages qui reposerait davantage sur des bases conceptuelles que statistiques. Une brève description de certains travaux qu'on y retrouve permettra peut-être de mieux saisir la diversité des réflexions et le bouillonnement d'idées auxquels les développements de la recherche en psychologie cognitive ont donné naissance.

En premier lieu, Frederiksen, Mislevy et Bejar (1993) présentent des travaux qui pourraient constituer (selon leurs propres mots) les fondements d'une nouvelle théorie des tests. Dans cette monographie, Lohman et Ippel y abordent (p. 59) la problématique du changement de stratégie des répondants lors du déroulement d'un test. L'individu s'adapte aux items et sa stratégie évolue selon le type de difficultés rencontrées. Après une analyse du processus de gestion de l'information, on peut proposer des modèles qui décrivent comment les répondants trouvent des solutions aux problèmes qui leurs sont proposés. Si on arrive à relier les stratégies utilisées à d'autres dimensions de l'apprentissage ou du développement cognitif, on pourrait proposer des pistes pour expliquer plus en profondeur les différences individuelles. Ensuite, Bennett y propose (p. 99) un modèle d'évaluation intelligente qui intègre des items à réponse construite, des procédures de codage des réponses s'inspirant de travaux en intelligence artificielle, et des modèles de mesure ayant des assises cognitives. De son côté, Marshall utilise (p. 155) le concept de schéma, développé en psychologie cognitive, pour disséquer en composantes simples les processus complexes. Elle produit des descriptions cognitives multidimensionnelles des répondants qui permettent d'aller au-delà de la réponse finale pour mieux comprendre le processus

menant à la production de cette réponse. Toutes les réflexions présentées dans cette monographie ont en commun de traiter le candidat, l'élève, comme une entité active qui interagit avec les items qui lui sont présentés. Ainsi, l'objectif central des recherches qui y sont présentées est d'analyser les réponses fournies par les candidats et d'essayer de comprendre comment cette activité conduit à une compréhension et à une résolution du problème soumis.

En deuxième lieu, Nichols, Chipman et Brennan (1995) présentent des travaux qui ont en commun de s'intéresser à une évaluation diagnostique cognitive. Parmi les préoccupations principales de cet ouvrage, on retrouve la modélisation du processus de réponse utilisé par l'élève et la construction de réseaux conceptuels. Dans cette monographie, Mislevy y décrit (p.43-71) une structure probabiliste visant à définir un espace de modèles de candidats. Ces modèles, qui sont des simplifications des réseaux d'habiletés, de connaissances et de stratégies, permettent de poser en fonction de probabilités les différences manifestées par les candidats dans les réponses à des questions ou lors de la résolution de problèmes. Il applique son approche à la soustraction de nombres entiers et de fractions.

En troisième lieu, Reynolds (1994) présente des travaux récents sur le sujet en ayant comme objectif une perspective multidisciplinaire sur les différences individuelles. On y retrouve la présentation de résultats de recherche en neuropsychologie, en psychologie cognitive et en psychométrie. Dans l'approche de Embretson, par exemple (pp. 107-136), le concepteur des items doit, en quelque sorte, devenir un expérimentateur mettant à l'épreuve la pertinence de tâches en fonction de leurs liens avec la théorie. Le développement des items devient un processus scientifique plutôt qu'intuitif. La gestion des items dans la banque se fait selon la complexité cognitive de ceux-ci et telle que précisée par les référents théoriques. Le lien est fait avec un modèle multidimensionnel de la théorie de la réponse à l'item en supposant que les paramètres d'un item représentent les différentes sources de complexité cognitive.

Évidemment, il n'y a pas de solution miracle qui suffirait pour toutes les situations, comme un nettoyant qui dissout toutes les taches. Il y a encore beaucoup de recherche à réaliser pour en arriver à élaborer des outils diagnostiques dont l'utilité pourrait être ressentie jusque dans la salle de classe. Mais, à la lumière des directions prises par les recherches sur le sujet, la prise de conscience nécessaire à une véritable révolution d'un *testing* plus directement associé à une évaluation des apprentissages qui vise le diagnostic est en bonne voie de se généraliser et de produire des résultats probants.

Conclusion

L'objectif de cet article n'était pas de nier les apports de la modélisation statistique. Pour sélectionner ou placer des individus dans des filières spéciales, on peut avancer qu'elle a fait ses preuves. La théorie classique des tests s'est développée en synergie avec la psychologie des grands groupes (Danziger, 1990); elle s'intéresse à des aggrégats de scores comme la moyenne ou la corrélation. C'est l'archétype d'un modèle de régression linéaire et c'est ce qui lui confère une certaine efficacité pour trier et sélectionner. Elle a été conçue avec ces objectifs en tête.

Un des avantages de la théorie de la réponse à l'item est de permettre d'isoler l'item du test. Ainsi, à partir d'une banque d'items ayant été précédemment étalonnés, on peut construire une quantité presque illimitée de tests qui seraient, sous certaines conditions, équivalents. Des applications au *testing* adaptatif laissent aussi entrevoir qu'il ne serait pas nécessaire que ces tests aient le même nombre d'items. Encore une fois, si l'objectif est de sélectionner, ce type de modèle devrait être assez efficace et, en plus, il serait la plupart du temps beaucoup plus économique. Si on se limite à cet objectif, les différences entre la théorie classique et la théorie de la réponse à l'item sont de l'ordre de l'efficacité statistique. On n'a pas changé la façon de corriger les items ou même, à la limite, de construire les tests. On répond aux mêmes impératifs tant que la modélisation statistique devance la modélisation conceptuelle.

Cependant, si on examine les travaux récents de Mislevy (1993), de Tatsuoka (1995) ou d'Embretson (1994), on s'aperçoit que la modélisation statistique change de rôle. Elle sévit en arrière-plan, au service de la conceptualisation et non l'inverse. En se rapprochant des processus de pensée, on se rapproche de la possibilité d'établir des diagnostics utiles pour l'élève. Au-delà de la simple prédiction des échecs ou des succès, on pourrait ainsi appliquer des modèles qui s'inscriraient dans une démarche de recherche d'information sur ce qui fait que les élèves savent ce qu'ils savent et font ce qu'ils font.

NOTES

1. C'est à celui-ci que l'on doit la formulation standard : X=T+E.
2. Le titre du livre de Lord et Novick, *Statistical theories of mental test scores*, est d'ailleurs très éloquent à cet égard.
3. Une bonne introduction à la théorie de la généralisabilité est le texte de Feldt et Brennan (1989, pp. 127-140).
4. Il demande en fait que les discriminations soient égales à l'unité. Des indices de discriminations communs autres que l'unité peuvent toujours être ramenés à celle-ci par une transformation appropriée.

RÉFÉRENCES

Ackerman, T.A. & Spray, J.A. (1986). A general model for item dependency. Texte présenté à la rencontre annuelle de l'American Educational Research Association, San Francisco.

Anderson, J., Kearney, G.E. & Everett, A.V. (1968). An evaluation of Rasch's structural model for test items. British Journal of Mathematical and Statistical Psychology, 21, 231-238.

Baker, F.B. (1959). Univac scientific computer program for test scoring and item analysis. Behavorial Science, 4, 254-255.

Baker, F.B. (1977). Advances in item analysis. Review of Educationnal Research, 47, 151-178.

Blais, J.-G. & Ajar, D. (1992). Théorie des réponses aux items et modélisation : loin de la coupe au lèvres. Mesure et évaluation en éducation, 14, 5-18.

Bock, D.R. & Wood, R. (1971). Test theory. Annual Review of Psychology, 22, 193-224.

Boring, E.G. (1961). The beginning and growth of measurement in psychology. Isis, 52, 238-257.

Cliff, N. (1989). Ordinal consistency and ordinal true score. Psychometrika, 54, 75-91.

Cronbach, L.J., Gleser, G.C. & Rajaratnam, N. (1963). Theory of generalizability. A liberalization of reliability theory. British Journal of Mathematical and Statistical Psychology, 16, 137-173.

Danziger, K. (1987). Statistical method and the historical development of research practice in American psychology. In L. Krüger, G. Gigerenzer & M.S. Morgan (éds), The probalistic revolution, vol. 2 : Ideas in the sciences. Cambridge, Mass. : MIT Press.

Danziger, K (1990). Constructing the subject. Cambridge : Cambridge University Press.

Divgi, D.R. (1989). Reply to Andrich and Henning. Journal of Educational Measurement, 26, 295-299.

Douglass, F.M., Khavari, K.A. & Farber, P.D. (1979). A comparison of classical and latent trait item analysis procedures. Educational and Psychological Measurement, 39, 337-352.

Embretson, S. (1983). Construct validity : construct representation versus nomothetic span. Psychological Bulletin, 93, 179-197.

Embretson, S. (1994). Applications of cognitive design systems to test development. In C.R. Reynolds (éd.), Cognitive assessment. A multidisciplinary prespective (pp. 107-136). New York : Plenum Press.

Evans, B. & Waites, B. (1981). IQ and mental testing : An unnatural science and its social history. Atlantic Highlands, N.J. : Humanities Press.

Feldt, L.S. & Brennan, R.L. (1989). Reliability. In R.L. Linn (éd.), Educational Measurement (3e édition). New York : ACE/Macmillan.

Ferguson, G.A. (1942). Item selection by the constant process. Psychometrika, 7, 19-29.

Fischer, G.H. (1987). Applying the principles of specific objectivity and of generalizability to the measurement of change. Psychometrika, 52, 565-587.

Fournier, M. (1994). Une perspective « kuhnienne » sur le développement de la théorie des réponses aux items. Mémoire de maîtrise non publié. Faculté des sciences de l'éducation, Université de Montréal.

Frängsmyr, T., Heilbron, J. & Rider, R. (éds) (1990). The quantifying spirit in the eighteenth century. Berkeley, CA : University of California Press.

Frederiksen,N., Mislevy,R.J. & Bejar, I (éds) (1993). Test theory for a new generation of tests. Hillsdale, N.J. : Lawrence Erlbaum Associates.

Goldstein, H. & Wood, R. (1989). Five decades of item response modelling. British Journal of Mathematical and Statistical Psychology, 42, 139-167.

Gould, S.J. (1983). La mal-mesure de l'homme. Paris : Éditions Ramsay.

Guilford, J.P. (1936). Psychometric methods. New York : McGraw-Hill.

Gullicksen, H. (1950). Theory of mental tests. New York : John Wiley.

Hacking, I. (1987). Was there a probalistic revolution 1800-1930? In L. Krüger, L.J. Daston & M. Heidelberger (éds), The probalistic revolution, vol. I : Ideas in history. Cambridge, Mass. : MIT Press.

Hambleton, R.K. (1986). The changing conception of measurement : a commentary. Applied Psychological Measurement, 10, 415-421.

Hambleton, R.K. & Swaminathan, H. (1985). Item response theory, principles and applications. Boston : Kluwer-Nijhoff.

Holland, P.W. (1990). On the sampling theory foundations of item response theory models. Psychometrika, 55, 577-601.

Kendler, H.H. (1987). Historical foundations of modern psychology. Philadelphie : Temple University Press.

Kuhn, T.S. (1983). La structure des révolutions scientifiques. Paris : Flammarion.

Lawley, D.N. (1943). On problems connected with item selection and test construction. Proceedings of the Royal Society of Edimburgh, 61, 273-287.

Lazarsfeld, P.F. (1950). Measurement and prediction. In S.A. Stouffer (éd.) Studies in social psychology in world war II, vol.4. Princeton : Princeton University Press.

Lord, F.M. (1952). A theory of test scores. Psychometrika monograph, No. 7.

Lord, F.M. (1953). An application of confidence intervals and of maximum likelihood to the estimation of an examinee's ability. Psychometrika, 18, 57-76.

Lord, F.M. (1965). A strong true-score theory, with applications. Psychometrika, 30, 239-270.

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

Luce, R.D. & Tuckey, J.W. (1964). Simultaneous conjoint measurement : a new type of fundamental measurement. Journal of Mathematical Psychology, 1, 1-27.

Lumsden, J. (1976). Test theory. Annual Review of Psychology, 27, 251-280.

Magnusson, D. (1967). Test theory. New York : John Wiley.

Maxwell, A.E. (1959). Maximum likelihood estimates of item parameters using the logistic function. Psychometrika, 24, 221-227.

McDonald, R.P. (1986). Describing the elephant : structure and function in multivariate data. Psychometrika, 51, 513-534.

Meijer, R.R., Sijtma, K. & Smid, N.G. (1990). Theoretical and empirical comparison of the Mokken and the Rasch approach to IRT. Applied Psychological Measurement, 14, 283-298.

Michell, J. (1990). An introduction to the logic of psychological measurement. Hillsdale, N.J. : Lawrence Erlbaum Associates.

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. Hillsdale, N.J. : Lawrence Erlbaum Associates.

Nichols, P.D., Chipman, S.F. & Brennan, R.L. (éds) (1995). Cognitively diagnostic assessment. Hillsdale, N.J. : Lawrence Erlbaum Associates.

Quetelet, A. (1835). Sur l'homme et le développement de ses facultés ou essai de physique sociale. Paris : Bachelier. (Réédition de 1991 chez Arthème-Fayard, Paris.)

Ramsay, J.O. (1989). A comparison of three simple test theory models. Psychometrika, 54, 487-499.

Rasch, G. (1960). Probalistic models for some intelligence and attainment tests. Copenhague : Danmarks Paedogogiske Institut.

Reckase, M.D. (1989). Adaptive testing : the evolution of a good idea. Educational Measurement : Issues and Practice, 8(3), 11-15.

Reuchlin, M. (1989). Histoire de la psychologie (14e édition). Paris : Presses Universitaires de France.

Reynolds, C.R. (éd.) (1994). Cognitive assessment. A multidisciplinary prespective. New York : Plenum Press.

Schneider, W.H. (1992). After Binet : french intelligence testing, 1900-1950. Journal of the history of the behavioral sciences, 28, 111-132.

Slinde, J.A. & Linn, R.L. (1979). The Rasch model, objective measurement, equating, and robustness. Applied Psychological Measurement, 3, 437-452.

Tatsuoka, K.K. (1995). Architecture of knowledge structure and cognitive diagnosis : a statistical pattern recognition and classification approach. In P.D. Nichols, S.F. Chipman & R.L. Brennan (éds), Cognitively diagnostic assessment. Hillsdale, N.J. : Lawrence Erlbaum Associates.

Thomson, G.O.B. & Sharp, S. (1988). History of mental testing. In J.P. Keeves, Educational research, methodology and measurement. An international handbook. Oxford : Pergamon Press.

Thurstone, L.L. (1931). The reliability and validity of tests. Ann Arbor, Mich. : Edwards Bros.

Thurstone, L. L. (1935). The vectors of mind : multiple-factor analysis for the isolation of primary traits. Chicago, Ill. : University of Chicago Press.

Tinsley, H.E.A. & Dawis, R.V. (1977). Test-free person measurement with the Rasch simple logistic model. Applied Psychological Measurement, 1, 483-487.

Traub, R.E. & Lam, Y.R. (1985). Latent structure and item sampling models for testing. Annual Review of Psychology, 36, 19-48.

Tucker, L.R. (1946). Maximum validity of a test with equivalent items. Psychometrika, 11, 1-13.

Van der Linden, W.J. (1986). The changing conception of measurement in education and psychology. Applied Psychological Measurement, 10, 325-332.

Wainer, H. & Lewis, C. (1990). Toward a psychometrics for testlets. Journal of Educational Measurement, 27, 1-14.

Weiss, D.J. & Davison, M.L. (1981). Test theory and methods. Annual Review of Psychology, 32, 629-658.

Whitely, S.E. (1977). Models, meanings and misunderstandings : some issues in applying Rasch's theory. Journal of Educationnal Measurement, 14, 227-235.

Whitely, S.E. (1980). Latent trait models in the study of intelligence. Intelligence, 4, 97-132.

Wise, M.N. (éd.) (1995). The values of precision. Princeton : Princeton University Press.

Wood, R. (1978). Fitting the Rasch model – a heady tale. British Journal of Mathematical and Statistial Psychology, 31, 27-32.

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

Wright, B.D. (1984). Despair and hope for educational measurement. Contemporary Education Review, 3, 281-288.

Wright, B.D. (1988). Rasch measurement models. In J.P. Keeves (éd.), Educational research, methodology and measurement. An international handbook (pp. 286-291). Oxford : Pergamon Press.