Mesure et évaluation en éducation

Une comparaison empirique de modèles de la théorie classique, de la théorie de la généralisabilité et de la théorie des réponses aux items

Richard Bertrand

Volume 26, Number 1-2, 2003

Généralisabilité

URI: https://id.erudit.org/iderudit/1088241ar
DOI: https://doi.org/10.7202/1088241ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

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

Explore this journal

Cite this article

Bertrand, R. (2003). Une comparaison empirique de modèles de la théorie classique, de la théorie de la généralisabilité et de la théorie des réponses aux items. *Mesure et évaluation en éducation*, *26*(1-2), 75–89. https://doi.org/10.7202/1088241ar

Article abstract

The aim of this article is to compare empirically measurement models from classical test theory, generalizability theory and item response theory. First the basic concepts of these measurement theories are presented. Then, using an attitude questionnaire of nine items administered to 3 600 subjects, the results of the psychometrical analyses associated with each of the theories are compared. Most of the psychometrical observations associated with a measurement theory are reproduced by the other two approaches. These results suggest that, even if these theories seem conceptually different, they lead to the same conclusions about basic psychometrical observations.

Une comparaison empirique de modèles de la théorie classique, de la théorie de la généralisabilité et de la théorie des réponses aux items

Richard Bertrand

Directeur du CRIRES, Centre interuniversitaire, Université Laval, Québec

MOTS CLÉS : Théorie classique des tests, théorie de la généralisabilité, théorie des réponses aux items, modélisation mathématique, fidélité, erreur type de mesure, coefficient alpha de Cronbach, coefficient de généralisabilité, information, modèle logistique à deux paramètres, modèle gradué de Samejima

L'objectif de cet article est de comparer empiriquement les modèles de mesure émanant de la théorie classique, de la théorie de la généralisabilité et de la théorie des réponses aux items. Après avoir traité des concepts de base qui caractérisent chacune de ces théories, le présent texte, s'appuyant sur les résultats d'un questionnaire d'attitude de neuf items administré à 3 600 sujets, compare les résultats des analyses métrologiques propres aux théories visées. La plupart des observations de nature métrologique faites à l'aide des résultats d'une analyse propre à une des théories peuvent être reproduites par les autres approches. Ces résultats suggèrent que même si elles paraissent conceptuellement distinctes, ces trois théories de la mesure ne peuvent que révéler les mêmes observations métrologiques de base.

KEY WORDS : Classical test theory, generalizability theory, item response theory, mathematical modelling, reliability, standard error of measurement, Cronbach's alpha, generalizability coefficient, information, 2-parameter logistic model, Samejima's graded model

The aim of this article is to compare empirically measurement models from classical test theory, generalizability theory and item response theory. First the basic concepts of these measurement theories are presented. Then, using an attitude questionnaire of nine items administered to 3 600 subjects, the results of the psychometrical analyses associated with each of the theories are compared. Most of the psychometrical observations associated with a measurement theory are reproduced by the other two approaches. These results suggest that, even if these theories seem conceptually different, they lead to the same conclusions about basic psychometrical observations.

PALAVRAS CHAVE: Teoria clássica dos testes, teoria da generalizabilidade, teoria da resposta a itens, modelização matemática, fidelidade, erro padrão de medida, coeficiente alfa de Cronbach, coeficiente de generalizabilidade, informação, modelo logístico com dois parâmetros, modelo graduado de Samejima

O objectivo deste artigo é comparar empiricamente os modelos de medida que emanam da teoria clássica, da teoria da generalizabilidade e da teoria da resposta a itens. Após ter abordado os conceitos de base que caracterizam cada uma destas teorias, o presente texto, apoiando-se nos resultados de um questionário de atitudes com nove itens administrado a 3 600 sujeitos, compara os resultados das análises metrológicas inerentes às teorias visadas. A maior

parte das observações de natureza metrológica, efectuadas com o auxílio da análise apropriada a uma das teorias, podem ser reproduzidas pelas outras abordagens. Estes resultados sugerem que, apesar de parecerem conceptualmente diferentes, estas três teorias da medida revelam as mesmas observações metrológicas de base.

Introduction

Peu d'ouvrages présentent des comparaisons empiriques des modèles de mesure émanant à la fois de la théorie classique, de la théorie de la généralisabilité et de la théorie des réponses aux items (Bertrand & Blais, 2003). La plupart des manuels se concentrent sur quelques modèles de l'une ou l'autre de ces théories sans jamais vraiment se référer aux modèles des deux autres : on évoque alors le manque d'espace ou la volonté de traiter d'une seule théorie mais dans ses plus fins détails. C'est bien légitime. Certains textes (Brennan, 1983; Brennan, 2001; McArthur, 1987) pourtant comportent un paragraphe comparant la nature de ces théories : on y présente en effet la théorie de la généralisabilité comme un prolongement de la théorie classique, toutes les deux étant perçues comme des théories d'échantillonnage (des items, des moments, des correcteurs, etc.), alors que la théorie des réponses aux items est conçue comme une théorie d'échelonnage (tablant sur la construction d'une échelle de mesure commune aux items administrés et aux sujets examinés). Le présent texte propose une comparaison empirique de ces trois théories en prenant appui sur les données[1] du Questionnaire sur la lecture du projet PISA (Program for International Student Assessment) de l'OCDE[2]. Nous avons voulu nous assurer que ces trois approches, même si elles n'étaient pas identiques en nature, fournissaient des données consistantes lorsqu'elles étaient comparées entre elles, démarche prolongeant ainsi les efforts de Bertrand (2002) en ce sens.

Concepts de base des trois théories

Commençons par bien distinguer la nature de chacune de ces trois théories. Plusieurs manuels (Allen & Yen, 1979; Crocker & Algina, 1986; Laveault & Grégoire, 2000; Lord & Novick, 1968; Suen, 1990; Traub, 1994) ont déjà abordé la présentation des concepts de base de la théorie classique. L'équation de base du modèle classique est donnée par :

$$X = V + E$$

où X est le score observé d'un individu, V est le score vrai de cet individu et E est l'erreur de mesure.

L'équation de base signifie que le score observé X, selon le modèle classique, possède deux composantes additives : V et E. Puisque ni le score vrai ni l'erreur de mesure ne sont observables, il faut procéder autrement pour déterminer l'ampleur de l'erreur qu'on commet en utilisant ce modèle de mesure. On s'en remet souvent au concept de fidélité ρ^2_{xv} et à celui d'erreur type de mesure σ^2 pour y parvenir. La fidélité peut être vue comme une proportion de variance vraie σ^2 dans la variance totale (observée) σ^2_x.

$$\rho^2_{XV} = \sigma^2_V / \sigma^2_X$$

L'erreur type de mesure, par ailleurs, est directement proportionnelle à la variance totale et inversement proportionnelle à la fidélité

$$\sigma_E = \sqrt{\sigma^2_X}\sqrt{(1-\rho^2_{XV})}$$

Il est courant, pour un test de *n* items, d'estimer la fidélité par le coefficient alpha de Cronbach (α)

$$\alpha = \left[\frac{n}{n-1}\right]\left[1 - \frac{\left(\sum_{i=1}^{n} s_i^2\right)}{s_X^2}\right]$$

et l'erreur type de mesure par

$$s_E = s_X \sqrt{1-\alpha}$$

où s_i^2 se réfère à la variance de l'item i et s_x^2 renvoie à la variance du test X composé de *n* items.

S'agissant de la théorie de la généralisabilité, ce sont les travaux de Cronbach, Rajaratnam et Gleser (1963), de Cronbach, Gleser, Nanda et Rajaratnam (1972), puis ceux de Brennan (1983) et Brennan (2001), et enfin ceux de Cardinet et Tourneur (1985), de Cardinet, Tourneur et Allal (1976, 1981) et de Bain et Pini (1996) qui ont permis à cette théorie de posséder la stature qu'elle présente maintenant.

La théorie classique en effet ne parvenait pas à apprécier adéquatement l'erreur de mesure de dispositifs complexes où plus d'une source d'erreur de mesure était impliquée. Les modèles de la théorie de la généralisabilité permettront de déterminer, de quantifier, puis de contrôler les différentes sources d'erreur de mesure qui demeurent indifférenciées en théorie classique, en posant comme équation de base :

$$X = V + \begin{bmatrix} E_1 \\ E_2 \\ E_3 \\ \vdots \\ E_k \end{bmatrix}$$

L'erreur de mesure est maintenant éclatée en plusieurs sources d'erreur différentes : E_1, E_2, E_3,... E_k, ce qui permet de quantifier la part relative de chacune de ces sources d'erreur. Le calcul d'un coefficient de généralisabilité, pendant du coefficient de fidélité en théorie classique, permet d'en arriver à cette quantification.

Deux types de coefficient de généralisabilité peuvent alors être envisagés, chacun étant fonction de l'objectif de la mesure. Si une décision relative doit être prise (par ex. : Quel est le meilleur élève ?), nous parlons de coefficient de généralisabilité relatif, le coefficient σ^2_p représentant la variance de différenciation ou variance vraie et le coefficient $\sigma^2{}_{\delta_p}$ la variance d'erreur :

$$\rho^2_{\delta_p} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{\delta_p}}$$

Si, par contre, l'intérêt de la mesure porte surtout sur une décision absolue (par ex. : Qui passe le seuil de 60 % ?), c'est alors à un coefficient de généralisabilité absolu qu'il faut se référer :

$$\rho^2_{\Delta_p} = \frac{\sigma^2_p}{\sigma^2_p + \sigma^2_{\Delta_p}}$$

La racine carrée de la variance d'erreur, qu'elle soit relative ($\sigma^2_{\delta_p}$) ou absolue ($\sigma^2_{\Delta_p}$), peut alors être considérée comme erreur type de mesure dans le cadre de la théorie de la généralisabilité.

Comme on peut le constater, plusieurs des concepts de base en théorie de la généralisabilité sont parallèles à des concepts de la théorie classique. En outre, nous montrerons plus loin comment la procédure de l'analyse d'items, tant utilisée en théorie classique, peut être étendue en théorie de la généralisabilité à une procédure connue sous le nom d'analyse de facettes. C'est en ce sens que la théorie classique est considérée comme un cas particulier de la théorie de la généralisabilité. La différence principale réside dans la constitution de la variance d'erreur qui, en théorie classique, fait référence à une seule source d'erreur indifférenciée, alors qu'elle contient, en théorie de la généralisabilité, plusieurs sources d'erreur bien reconnaissables.

Les textes de Baker (1992), Hambleton et Swaminathan (1985), Hambleton, Swaminathan et Rogers (1991), Hulin, Drasgow et Parsons (1983), Lord (1980) ou Thissen et Wainer (2001) témoignent de l'importance grandissante de la théorie des réponses aux items. C'est à un changement de paradigme que nous invite cette théorie de la mesure si on la compare aux deux théories précédentes.

Il s'agit en fait de modéliser mathématiquement la rencontre entre un sujet à qui est administré un questionnaire (test, échelle, etc.) et un item (question, énoncé, etc.) de ce questionnaire. Puisqu'en général plusieurs modèles peuvent entrer en compétition pour rendre compte au mieux de la relation entre un sujet examiné et un item administré, la question de la qualité de l'ajustement du modèle aux données résultant de l'administration du questionnaire aux sujets prend ici une importance capitale. De même, le nombre de traits mesurés par ce questionnaire doit être réglé avant de procéder à l'application du modèle.

Les modèles les plus simples visent un seul trait et une vérification de l'unidimensionnalité doit donc être entreprise. Dans le cas d'un modèle unidimensionnel, la représentation graphique, appelée courbe caractéristique d'item (CCI), de la rencontre entre les sujets examinés et un item du questionnaire peut prendre l'allure de la figure 1.

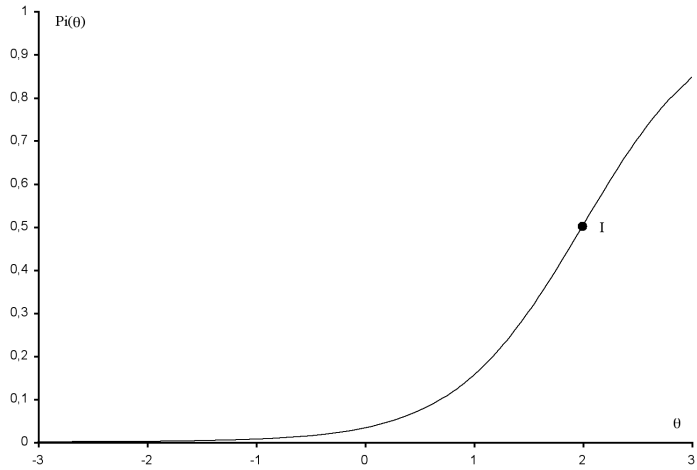

Figure 1. ***Courbe caractéristique d'item exprimant la relation entre, en abscisse, l'habileté d'un sujet (θ) et, en ordonnée, la probabilité (Pi (θ)) de réussir (endosser) cet item.***

La représentation algébrique du modèle est donnée par :

$$P_i(\theta) = \frac{1}{1+e^{-Da_i(\theta-b_i)}}$$

où θ indique l'habileté (capacité, performance, attitude, etc.) d'un sujet examiné, P_i (θ) exprime la probabilité de réussir (endosser) l'item i pour le sujet d'habileté θ, b_i est un indice de difficulté (de réussir, endosser) l'item i et a_i est un indice de la capacité de discriminer de l'item i.

L'échelle d'habileté permet de comparer l'habileté θ_j d'un sujet j,avec la difficulté, b_i, d'un item i. Plus le sujet obtient une valeur élevée (généralement entre –3 et +3) sur l'échelle d'habileté θ et plus le sujet est considéré comme ayant une grande capacité en regard de l'habileté visée. Puisque b_i est l'abscisse du point d'inflexion I, alors plus un item obtient une valeur b_i élevée, c'est-à-dire plus la CCI est déplacée vers la droite, et plus l'item est considéré difficile.

L'indice de discrimination de l'item i, a_i, est proportionnel à la pente de la tangente au point d'inflexion I. Ainsi, plus la pente de la CCI est élevée au point d'inflexion et plus l'item discrimine dans le voisinage du point d'inflexion, c'est-à-dire dans le voisinage de b_i.

Le concept d'information (Bertrand & Blais, 2003) est considéré comme le pendant du concept de fidélité en théorie classique et du concept de généralisabilité dans la théorie correspondante. Contrairement à ces deux théories cependant on peut obtenir, en théorie des réponses aux items, une valeur d'information pour chaque item et à chaque valeur de l'échelle d'habileté θ. La valeur de l'information au point de l'échelle θ est proportionnelle à la pente de la CCI au point θ.

Revenant à la figure 1, on voit, par exemple, que la probabilité de réussir (endosser) cet item pour un sujet d'habileté moyenne ($\theta = 0$) est très faible : environ 0,05. Cette probabilité augmente à mesure que l'habileté augmente à tel point qu'un sujet d'habileté élevée ($\theta = 2$) a près de 50 % de chance de réussir cet item et qu'un sujet encore plus habile ($\theta = 3$) a plus de 80 % de chance de le réussir. La valeur de l'indice de difficulté b_i de cet item est égale à 2.

On peut observer en outre que l'intervalle de l'échelle d'habileté θ où l'item discrimine le plus, donc où l'item donne le plus d'information, se situe de part et d'autre de la valeur de l'indice de difficulté, soit dans le voisinage de $b_i = 2$.

Comparaisons empiriques des trois théories

C'est grâce à un exemple proposé par Daniel Bain que nous allons pouvoir comparer les résultats des analyses d'un questionnaire d'attitude face à la lecture à l'aide des trois théories de la mesure. Le questionnaire comprend en tout neuf items et les données de 3 600 répondants proviennent de l'enquête PISA de l'OCDE administrée en 2000 en Suisse romande. Quatre catégories de l'échelle de mesure ont été utilisées : pas du tout d'accord, pas d'accord, d'accord, et tout à fait d'accord.

Nous allons traiter les données de ce questionnaire d'attitude à l'aide de quatre modèles de mesure : le modèle classique, un modèle de généralisabilité orienté vers une décision relative, le modèle logistique de réponses aux items à deux paramètres décrit plus haut et le modèle gradué de Samejima, que nous décrirons plus loin.

Le tableau 1 présente les résultats de l'analyse classique. Les valeurs sous la colonne titrée « Moyenne » indiquent le score moyen[3] obtenu par les 3 600 sujets à chacun des neuf items. On voit également à ce tableau que les valeurs de l'indice de discrimination, c'est-à-dire les valeurs de la corrélation item-total (corrigée), sont très élevées, indiquant que les neuf items sont intimement liés ensemble, du moins empiriquement.

La valeur du coefficient alpha est également très élevée, signifiant une très grande consistance interne de cette échelle de neuf items. Remarquons encore que c'est l'item 1 (« La lecture est un de mes loisirs favoris ») qui discrimine le plus et l'item 9 (« J'éprouve des difficultés à finir les livres »[4]) qui discrimine le moins. D'ailleurs, c'est seulement en enlevant ce dernier item que la valeur du coefficient alpha s'améliore, passant de 0,905 à 0,906.

Enfin, nous avons calculé la valeur du coefficient de corrélation intraclasse, ici le coefficient de Spearman-Brown, dans le cas où on suppose le questionnaire réduit à un seul item : nous avons trouvé une valeur de 0,515.

Tableau 1
Résultats de l'analyse classique des neuf items du questionnaire d'attitude selon SPSS 11.0

Numéro de l'item	*Moyenne*	*Corrélation item-total*	*Alpha si l'item est enlevé*
1	2,169	0,771	0,888
2	2,119	0,650	0,897
3	2,438	0,695	0,893
4	2,464	0,657	0,896
5	2,937	0,769	0,888
6	3,061	0,759	0,889
7	2,601	0,635	0,898
8	2,992	0,662	0,896
9	2,919	0,516	0,906
Coefficient alpha = 0,905		Coefficient de Spearman-Brown (k=1) = 0,515	

Le tableau 2 présente les résultats de l'analyse de généralisabilité (relative) où deux facettes aléatoires infinies croisées sont impliquées : une facette de différenciation, les élèves (E), et une facette d'instrumentation, les items (I).

On note tout d'abord que la valeur du coefficient de généralisabilité est de 0,905 soit la même valeur que le coefficient alpha trouvé au tableau 1. Rien d'étonnant : le coefficient de généralisabilité relatif dans le cas de deux facettes croisées aléatoires infinies est très exactement égal au coefficient alpha de Cronbach.

En pratiquant une optimisation à rebours, on se rend compte que le coefficient de généralisabilité relatif passe à 0,515 en supposant la généralisabilité basée sur un seul item. Il s'agit en réalité de la même valeur que celle du coefficient de Spearman-Brown relevée au tableau 1.

Par ailleurs, l'analyse de la facette I donne la même information que l'analyse d'items du tableau 1 : la valeur du coefficient alpha, si l'item est enlevé, est la même que la valeur du coefficient de généralisabilité relatif si le niveau donné de la facette I est enlevé.

Tableau 2
Résultats de l'analyse de généralisabilité des items du questionnaire d'attitude selon EduG 1.8

Plan d'observation et d'estimation			*Analyse de la facette I*	
Facette	*Niveaux observ.*	*Univers*	*Sans ce niveau*	*Fidélité relat.*
E	3 600	INF	1	0,888
I	9	INF	2	0,897
			3	0,893
Coefficients de généralisabilité			4	0,896
Relatif	0,905		5	0,888
Absolu	0,882		6	0,899
			7	0,898
Optimisation			*8*	*0,896*
Avec un seul item		0,515	9	0,906

Ces dernières observations nous montrent jusqu'à quel point la théorie de la généralisabilité peut être vue comme une extension bien naturelle de la théorie classique, non seulement pour ce qui est de l'extension des concepts, mais aussi de l'extension des méthodes.

Qu'ont maintenant en commun ces deux théories et la théorie des réponses aux items ? Afin de répondre à cette question, nous avons choisi d'analyser les données du questionnaire d'attitude à l'aide de deux modèles de réponses aux items. Le premier est dichotomique, c'est-à-dire qu'il s'appuie sur les données dichotomisées de l'échelle de mesure : nous donnons donc une valeur de 0 aux catégories « pas du tout d'accord » et « pas d'accord » et une valeur de 1 aux catégories « d'accord » et « tout à fait d'accord ».

Il s'agit du modèle logistique à deux paramètres. Le deuxième modèle est polytomique ordinal et prendra appui sur les quatre catégories ordinales pour les analyses. Il s'agit en fait du modèle gradué de Samejima (Bertrand & Blais, 2003 ; Thissen & Wainer, 2001).

Les figures 2 et 3 présentent les courbes caractéristiques[5] respectives des items 1 et 9 du questionnaire d'attitude.

Ce qui frappe tout d'abord en comparant ces deux CCI, c'est la différence de l'indice de difficulté ($b_1 = 0,44$; $b_9 = -0,80$) : cela montre que l'item 1 est plus difficile à endosser que l'item 9. Cette observation est corroborée par les

valeurs de l'indice classique de difficulté, noté «Moyenne» au tableau 1 : 2,169 pour l'item 1 et 2,919 pour l'item 9. Pas étonnant, puisqu'il paraît objectivement plus difficile d'endosser l'item 1 («La lecture est un de mes loisirs favoris») que l'item 9 («Je n'éprouve pas de difficulté à finir les livres»[6]).

L'item 1 est par ailleurs plus discriminant que l'item 9 : $a_1 = 2{,}670$ et $a_9 = 1{,}380$. On peut le constater en examinant la pente de la CCI : celle de l'item 1 est plus abrupte que celle de l'item 9 dans le voisinage du point d'inflexion. Observation corroborée au tableau 1 par les valeurs de l'indice de discrimination, la corrélation item-total, des items 1 ($r_{it1} = 0{,}770$) et 9 ($r_{it9} = 0{,}516$). On pourra noter aussi que la valeur de l'information maximale de l'item 1 est de 0,6 au point d'habileté $\theta = b_1 = 0{,}44$ alors qu'elle est d'à peine 0,15 au point d'habileté $\theta = b_9 = -\ 0{,}80$ pour l'item 9.

La valeur de l'information maximale pour l'ensemble du questionnaire est de 10 : l'erreur type de mesure[7] est donc de 0,316.

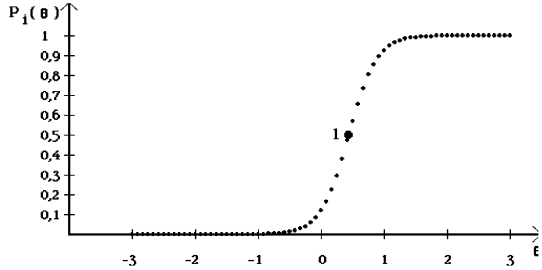

Figure 2. ***Item 1 de l'échelle d'attitude $a_1 = 2{,}670$ et $b_1 = 0{,}44$; modèle logistique à deux paramètres selon Bilog 3 (Mislevy & Bock, 1996); l'information maximale est de 0,6 au point $b_1 = 0{,}44$.***

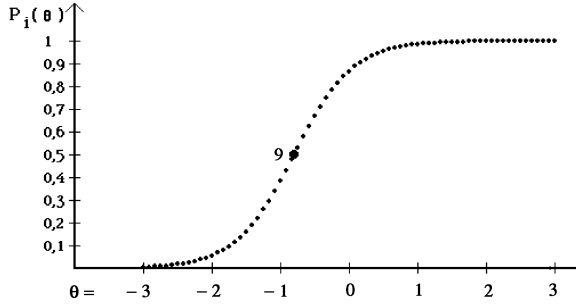

Figure 3. ***Item 9 de l'échelle d'attitude $a_9 = 1{,}380$ et $b_9 = -\ 0{,}80$; modèle logistique à deux paramètres selon Bilog 3 (Mislevy & Bock, 1996); l'information maximale est de 0,15 au point $b_9 = -\ 0{,}80$.***

Les figures 4 et 5 révèlent les courbes caractéristiques des items 1 et 9 du questionnaire d'attitude face à la lecture selon le modèle gradué de Samejima.

On remarque qu'il y a une courbe pour chacune des quatre catégories de l'échelle de mesure : la courbe notée 1 représente la catégorie « pas du tout d'accord », la courbe 2 la catégorie « pas d'accord », la courbe 3 « d'accord » et la courbe 4 « tout à fait d'accord ». Il est bien normal que la courbe 4 se situe à droite de la courbe 1 puisqu'il faut avoir une habileté θ plus grande, c'est-à-dire avoir une attitude plus positive envers la lecture, pour choisir la catégorie « tout à fait d'accord » plutôt que la catégorie « pas du tout d'accord » à un item comme « La lecture est un de mes loisirs favoris ».

Remarquons encore que la courbe 4 de la figure 4 comporte une ressemblance certaine avec la CCI de la figure 2 et que la courbe 4 de la figure 5 ressemble également à la CCI de la figure 3. Il faut aussi noter que les pentes des courbes sont plus abruptes à la figure 4 (item 1) qu'à la figure 5 (item 9), conformément à ce que nous avions observé aux figures 2 et 3, car l'item 1 discrimine plus que l'item 9. D'ailleurs, l'information maximale de l'item 1, selon le modèle de Samejima est de 2,4 alors qu'elle est de 0,4 pour l'item 9.

Remarquons aussi que les courbes de la figure 4 sont décalées vers la droite si on les compare à celles de la figure 5, signifiant que l'item 1 est globalement plus difficile à endosser que l'item 9, une observation que nous avions déjà faite.

Notons enfin que l'information (précision) maximale du questionnaire est de 11 dans le cas du modèle gradué de Samejima, ce qui constitue une légère amélioration par rapport à l'information du questionnaire obtenue selon le modèle à deux paramètres puisque la valeur de l'information maximale était de 10.

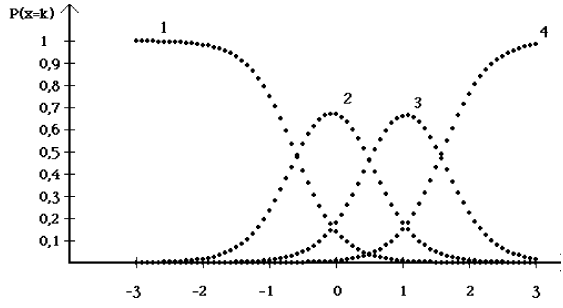

Figure 4. ***Item 1 de l'échelle d'attitude a = 2,909 b_1= – 0,618 b_2 = 0,501 b_3 = 1,599; modèle gradué de Samejima selon Multilog (Thissen, 1991); l'information maximale est de 2,4.***

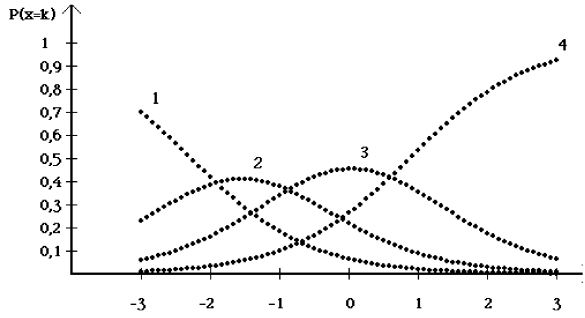

Figure 5. ***Item 9 de l'échelle d'attitude a = 1,173;b_1= – 2,282 ; b_2 = – 0,790; b_3 = 0,882; modèle gradué de Samejima selon Multilog (Thissen, 1991); l'information maximale est de 0,4.***

Conclusion

Cet article a présenté une comparaison empirique de quatre modèles de mesure émanant de trois théories conceptuellement distinctes : la théorie classique, la théorie de la généralisabilité et la théorie des réponses aux items.

Nous avons montré, par l'analyse d'un questionnaire de neuf items d'attitude face à la lecture, que ces théories ne se contredisaient pas, loin de là. En réalité, la plupart des caractéristiques des items (difficulté, discrimination) ou, plus globalement du questionnaire (fidélité, généralisabilité, information) pouvaient être corroborées chez l'un ou l'autre des modèles de mesure.

Ces observations nous mènent naturellement à nous demander pourquoi utiliser un autre modèle que le modèle classique, si connu, si simple.

- Tout d'abord, s'agissant de la théorie de la généralisabilité, il n'existe pas de théorie concurrente pouvant traiter de plusieurs facettes à la fois : le modèle classique peut être comparé avec un modèle de généralisabilité, mais seulement dans le cas où deux facettes croisées aléatoires infinies sont en cause.
- Par ailleurs, les modèles de la théorie des réponses aux items, s'ils sont congruents avec le modèle classique, donnent des indices beaucoup plus précis et permettent des applications très difficilement envisageables dans un contexte classique. Par exemple, en TRI, les valeurs d'information et

donc d'erreur type de mesure peuvent être évaluées à chaque niveau d'habileté : une situation peu conforme avec le modèle classique où une valeur unique d'erreur type de mesure est le plus souvent[8] rapportée.

- Enfin, des applications comme le *testing* adaptatif par ordinateur (Bertrand, 2001) ne peuvent définitivement pas être envisagées sans le recours aux modèles de la théorie des réponses aux items.

NOTES

1. Données gracieusement mises à disposition par le Consortium PISA de Suisse romande (Chr. Nidegger, SRED), par l'intermédiaire de Daniel Bain.
2. Bertrand, R. (2002). *Le logiciel EduG pour les études de généralisabilité : quelques repères*. Communication présentée dans le cadre du Congrès de l'ADMEE-Europe. Lausanne, septembre 2002.
3. En tenant compte de l'inversion de l'échelle pour certains items, une valeur de 1 est attribuée à la catégorie « pas du tout d'accord » et une valeur de 4 à la catégorie « tout à fait d'accord ».
4. L'échelle de mesure de cet item a bien sûr été inversée.
5. L'examen du graphique des éboulis (« scree plot ») nous a convaincu qu'une dimension dominante était présente : la première valeur propre étant de cinq fois plus élevée que la seconde valeur propre. De même, l'examen des courbes caractéristiques d'item produites par Bilog3 a révélé un ajustement presque parfait du modèle logistique à deux paramètres.
6. Interprétation libre de l'inversion de l'item original « J'éprouve des difficultés à finir les livres ».
7. L'erreur type de mesure en TRI est donnée comme l'inverse de la racine carrée de l'information.
8. Woodruff (1990) a proposé le calcul d'une valeur d'erreur type de mesure à chaque niveau d'habileté, dans un contexte d'analyse classique, mais pour y arriver, il faut compter sur plusieurs centaines de sujets !

RÉFÉRENCES

Allen, M.J. & Yen, W.M. (1979). *Introduction to measurement theory*. Monterey : Brooks & Cole.

Baker, F.B. (1992). *Item response theory : parameter estimation techniques*. New York, NY : Marcel Dekker.

Bain, D. & Pini, G. (1996). *Pour évaluer vos évaluations. La généralisabilité : mode d'emploi*. Genève : Centre de recherches psychopédagogiques. Direction générale du Cycle d'orientation.

Bertrand, R. (1994). CASANOVA : une méthode graphique d'identification des composantes de variance. In L. Laurencelle (éd.), *Trois essais en méthodes quantitatives*. Sillery : Presses de l'Université du Québec.

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

Bertrand, R. (2002). *Le logiciel EduG pour les études de généralisabilité : quelques repères*. Communication dans le cadre du Symposium *La généralisabilité : un instrument pour tester la qualité des dispositifs d'évaluation* (J. Cardinet). XV[e] Colloque international de l'ADMEE-Europe, septembre 2002, Lausanne.

Bertrand, R. & Blais, J.G. (à paraître). *Modèles de mesure : l'apport de la théorie des réponses aux items*. Québec : Presses de l'Université du Québec.

Brennan, R.L. (1983). *Elements of generalizability theory*. Iowa City : The American College Testing Program.

Brennan, R.L. (2001). *Generalizability theory*. New York : Springer-Verlag.

Cardinet, J. & Tourneur, Y. (1985). *Assurer la mesure*. Berne : Peter Lang.

Cardinet, J., Tourneur, Y. & Allal, L. (1976).The symmetry of generalizability theory : Applications to educational measurement. *Journal of Educational Measurement*, *13*, 119-135.

Cardinet, J., Tourneur, Y. & Allal, L. (1981). Extension of generalizability theory and its applications in educational measurement. *Journal of Educational Measurement*, *18*, 183-204, et *19*, 331-332.

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York : Holt, Rinehart & Winston.

Cronbach, L.J., Rajaratnam, N. & Gleser, G.C. (1963). Theory of generalizability : A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*, 137-163.

Cronbach, L.J., Gleser, G.C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements : Theory of generalizability for scores and profiles*. New York : Wiley.

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

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

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

Laveault, D. & Grégoire, J. (2002). *Introduction aux théories des tests en psychologie et en sciences de l'éducation (2[e] éd.)*. Bruxelles : De Boeck.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ : Lawrence Erlbaum.

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

McArthur, D.L. (éd.) (1987). *Alternative approaches to the assessment of achievement*. Boston, MA : Kluwer-Nijhoff Publishing.

Mislevy, R.J. & Bock, R.D. (1996). *BILOG-WINDOWS : Item analysis and test scoring with binary logistic models*. Mooresville : Scientific Software Inc.

Shavelson, R.J. & Webb, N.M. (1991). *Generelizability theory : a primer.* Newbury Park, Ca : Sage Publications.

Suen, H.K. (1990). *Principles of test theories*. Hillsdale : Lawrence Erlbaum.

Thissen, D. (1991). *MULTILOG user's guide : multiple, categorical item analysis and test scoring using item response theory*. Chicago, IL : Scientific Software.

Thissen, D. & Wainer, H. (2001). *Test scoring*. Mahwah, NJ : Lawrence Erlbaum.

Traub, R.E. (1994). *Reliability for the social sciences.* Newbury Park : Sage Publications.

Woodruff, D. (1990). Conditional standard error of measurement in prediction. *Journal of Educational Measurement*, *27*, 191-208.