Mesure et évaluation en éducation

Influence des distributions du trait latent et de la difficulté des items sur les estimations du modèle de Birnbaum : une étude du type Monte-Carlo

Réginald Burton

Volume 27, Number 3, 2004

URI: https://id.erudit.org/iderudit/1087787ar
DOI: https://doi.org/10.7202/1087787ar

See table of contents

Publisher(s)

ADMEE-Canada - Université Laval

ISSN

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

Explore this journal

Cite this article

Burton, R. (2004). Influence des distributions du trait latent et de la difficulté des items sur les estimations du modèle de Birnbaum : une étude du type Monte-Carlo. *Mesure et évaluation en éducation*, *27*(3), 41–62. https://doi.org/10.7202/1087787ar

Article abstract

IRT [item response theory] models require validity conditions which turn out to limit their applications. The IRT technicity, the stakes they are aimed at have not always given the opportunity of argued debates about their limits. However, as far as orientation or certification procedures are concerned, the precision level of evaluations effectuated on the basis of IRT models is very important since, in those aspects, students results can have heavy consequences for their future. That is the reason why this study will try to determine which are the features required for a sample of students or of items in order to secure the quality level of Birnbaum's model estimates.

Influence des distributions du trait latent et de la difficulté des items sur les estimations du modele de Birnbaum : une étude du type Monte-Carlo

Réginald Burton
Université du Luxembourg

MOTS CLÉS : Modèle de réponse à l'item, estimation, évaluation

« Les MRI [modèles de réponse à l'item] requièrent des conditions de validité qui aboutissent à limiter leurs champs d'application. Or, la technicité des MRI, les enjeux dont ils sont l'objet, n'ont pas toujours permis une discussion argumentée de leurs limites » (Vrignaud & Chartier,1999). Cependant, dans le cadre de procédures d'orientation ou de certification, le degré de précision des estimations réalisées au départ des MRI revêt un caractère particulièrement important puisque, dans ces cas, les résultats des élèves peuvent avoir des conséquences graves sur leur avenir. Aussi tenterons-nous de déterminer, dans cette étude, quelles sont les caractéristiques que doit remplir un échantillon d'élèves et d'items pour garantir la qualité des estimations des paramètres du modèle de Birnbaum.

KEY WORDS : Item response theory, estimation, assessment

IRT [item response theory] models require validity conditions which turn out to limit their applications. The IRT technicity, the stakes they are aimed at have not always given the opportunity of argued debates about their limits. However, as far as orientation or certification procedures are concerned, the precision level of evaluations effectuated on the basis of IRT models is very important since, in those aspects, students results can have heavy consequences for their future. That is the reason why this study will try to determine which are the features required for a sample of students or of items in order to secure the quality level of Birnbaum's model estimates.

PALAVRAS-CHAVE : Modelo de resposta ao item, estimação, avaliação

« Os MRI [modelos de resposta ao item] requerem condições de validade que conduzem a limitar os seus campos de aplicação. Ora, a tecnicidade dos MRI, os desafios de que são objecto, nem sempre têm permitido uma discussão argumentada dos seus limites » (Vrignaud & Chartier, 1999). Entretanto, no quadro de procedimentos de orientação ou de certificação, o grau de precisão das

estimações realizadas, relativas aos MRI, assume um carácter importante uma vez que, nestes casos, os resultados dos alunos podem ter consequências graves no seu futuro. Assim, tentaremos determinar, neste estudo, quais são as características que deve ter uma amostra de alunos e de itens para garantir a qualidade das estimações dos parâmetros do modelo de Birnbaum.

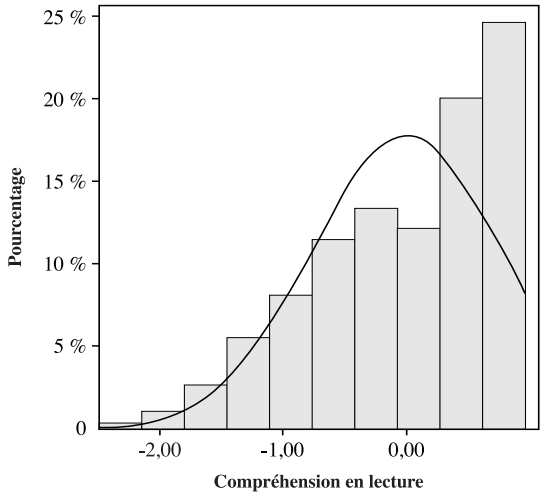
Introduction

Comme l'attestent Vrignaud et Chartier (1999), l'utilisation des modèles de réponse à l'item (MRI) s'est largement généralisée depuis une vingtaine d'années dans les comparaisons longitudinales et internationales portant sur les performances des élèves et l'évaluation des politiques d'éducation. Ils sont employés notamment par les deux plus grands organismes chargés de l'évaluation des acquis des élèves sur le plan international : l'IEA (International Association for the Evaluation of Educational Achievement) et l'OCDE (Organisation de coopération et de développement économiques). Les MRI ont apporté des solutions élégantes à des questions telles que la construction d'une variable latente permettant la comparaison de sujets n'ayant pas passé des épreuves identiques ou l'identification des biais de réponse entre populations (fonctionnement différentiel des items). Cependant, les MRI requièrent des conditions de validité qui aboutissent à limiter leurs champs d'application. Mais, la technicité des MRI, les enjeux dont ils sont l'objet, n'ont pas toujours permis une discussion argumentée de leurs limites.

Et bien que, dans le cas du modèle de Rasch (1980), il existe des méthodes d'estimation robustes même si le trait latent ne se distribue pas normalement (Follmann, 1988), la plupart des études réalisées jusqu'à présent, sur les MRI à deux ou trois paramètres et leurs méthodes d'ajustement, supposent que la compétence des sujets qui fait l'objet de l'évaluation est distribuée normalement au sein de la population testée (Baker, 1992 ; Habermann, 1977).

Or, il est de nombreux cas en sciences de l'éducation où les compétences évaluées sont censées être maîtrisées par une proportion importante des sujets auxquels on fait passer les tests. Cela peut se produire notamment lorsqu'on désire évaluer les compétences minimales d'élèves au terme d'un apprentissage. Dans ce cas, les compétences des sujets ne sont plus distribuées normalement mais selon des distributions asymétriques dont le mode est d'autant décalé vers la droite que la compétence des sujets est grande dans le domaine évalué. Inversement, avant tout apprentissage, il existe des situations où la distribution de l'aptitude des sujets est faible et présente une asymétrie positive.

À titre d'exemple, lors du passage de l'école primaire à l'école secondaire, les élèves de sixième année primaire du Grand-Duché du Luxembourg sont soumis à une évaluation de leurs compétences en mathématiques, en français et en allemand, dont les résultats conditionnent leur orientation scolaire. Des MRI sont utilisés pour estimer les aptitudes des élèves au départ de leurs réponses à un ensemble d'items conçus pour les besoins de l'évaluation. Cependant, pour nombre de compétences évaluées telles que la compréhension en lecture en allemand, il apparaît clairement que la distribution des aptitudes n'est pas normale mais en forme de «J», traduisant ainsi la maîtrise du domaine considéré pour une proportion importante d'individus. Le graphique suivant montre distinctement que la distribution de la compétence des élèves de sixième année primaire en compréhension en lecture en allemand n'est pas symétrique. Le test de Kolmogorov-Smirnov appliqué à ces données atteste d'ailleurs de la non-normalité de la distribution, avec une probabilité inférieure à 0,001. Notons, par ailleurs, que la distribution est caractérisée par une faible dispersion puisque l'écart type qui y est associé est égal à 0,78 et par une asymétrie négative prononcée (skewness = -0,91).



Graphique 1. ***Distribution de la compétence des élèves luxembourgeois de sixième primaire en compréhension en lecture en allemand***

Source : Épreuves standardisées 2000/2001. Documentation des résultats des épreuves standardisées (ministère de l'Éducation nationale, de la Formation professionnelle et des Sports)

Un autre problème qui peut également survenir dans de telles évaluations consiste à estimer la compétence des sujets au départ d'un ensemble d'items dont les caractéristiques psychométriques sont relativement homogènes. Ainsi, dans notre exemple, les neuf items qui sont utilisés pour évaluer la compréhension globale de textes présentent des indices de difficulté relativement peu dispersés puisque, avec une moyenne –1,490 et un écart type de 0,877, ils varient de -3,251 à -0,226. Il serait donc légitime de s'interroger sur l'opportunité d'employer de tels modèles dans ce type de situations, d'autant plus que la variance de la distribution de l'aptitude des sujets de l'échantillon joue un rôle prépondérant dans la stabilité des estimations des paramètres. Ainsi, un échantillon homogène de sujets peut entraîner des estimations instables des paramètres du modèle (Hambleton, 1994).

Aussi tenterons-nous de déterminer, dans cette étude, quelles sont les caractéristiques psychométriques qu'une épreuve d'évaluation doit remplir pour garantir la qualité des estimations de la compétence des sujets, dans le cas où le trait latent est distribué de manière asymétrique dans la population et que les estimations sont réalisées par l'intermédiaire du modèle de Birnbaum.

Méthode

Afin d'analyser l'influence de la distribution d'échantillonnage de la compétence des sujets et de la difficulté des items sur la qualité des estimations des paramètres du modèle de Birnbaum (1968), nous avons réalisé une étude du type Monte-Carlo, qui s'effectue en trois temps.

Dans un premier temps, des matrices de réponses sont générées à partir d'échantillons aléatoires et simples d'items et de sujets. Les échantillons de sujets sont issus de populations-parentes caractérisées par la nature, le degré d'asymétrie et l'ampleur de la dispersion de la distribution de leur habileté. Selon les objectifs poursuivis, deux types de distributions sont considérés : des distributions Bêta, qui présentent une asymétrie négative pour simuler des échantillons de sujets dont la compétence est élevée, et des distributions normales centrées réduites qui serviront de témoin à la comparaison.

Dans un deuxième temps, la compétence des sujets et la difficulté des items sont estimées dans le cadre du modèle de Birnbaum, à partir des matrices de réponses simulées. Les estimations sont réalisées par l'intermédiaire du logiciel BILOGMG, qui fait appel à la méthode du maximum de vraisemblance marginale pour l'estimation des paramètres des items (Zimowski, Muraki, Mislevy & Bock, 1996).

Dans un troisième temps, nous avons appréhendé la qualité des estimations issues du logiciel BILOGMG quant à leur précision. Dans cette perspective, nous avons comparé les estimations des paramètres aux valeurs théoriques qui ont servi à la construction des matrices de réponses.

Ainsi, dans les paragraphes qui suivent, après avoir défini les paramètres envisagés dans notre étude, nous décrirons les générateurs de nombres pseudo-aléatoires utilisés pour simuler la compétence des sujets et les paramètres des items. Nous décrirons également comment les matrices de réponses des sujets aux items ont été générées. Et, en fonction du domaine de définition des paramètres de l'étude, nous établirons le plan de simulation. Pour finir, nous exposerons la manière dont nous avons envisagé l'estimation des paramètres et le traitement des résultats.

Définition des paramètres

Trois paramètres seront pris en compte dans la simulation :

- le type de distribution du trait latent dans la population caractérisée par la nature, le degré d'asymétrie et la dispersion,
- le degré d'homogénéité de la distribution de la difficulté des items,
- le nombre d'items servant à évaluer la compétence des sujets.

Les choix des modalités des paramètres, qui sont décrits ci-dessous, ont été effectués en fonction des conditions que l'on rencontre habituellement dans la procédure d'orientation des élèves de sixième primaire au Grand-Duché du Luxembourg.

Le premier facteur envisagé dans cette étude concerne donc la distribution du trait latent dans la population, qui sera caractérisée par trois paramètres : la nature, le degré d'asymétrie et l'ampleur de la dispersion. Nous en avons considéré deux types (tableau 1) :

- des distributions Bêta (asymétrie négative, forme de J) pour simuler des échantillons de sujets dont la compétence est élevée,
- une distribution normale centrée réduite qui servira de témoin à la comparaison.

En considérant la distribution normale réduite comme référence, la dispersion de la distribution Bêta a été fixée à une valeur de moindre importance pour rendre compte de son homogénéité relative. Un écart type de 0,75 a été choisi.

Le second facteur se rapporte à la distribution des paramètres de difficulté des items et plus précisément à leur degré d'homogénéité. Dans cette perspective, deux distributions uniformes respectivement dans les intervalles [-3 ; 3] et [-3 ; -1] ont été utilisées.

Tableau 1
Caractéristiques des distributions utilisées pour simuler la compétence des sujets

N°	*Nom*	*Type*	*Asymétrie (γ_1)*	*Caractéristiques*		*Moyenne*	*Écart type*
1	n	normale	0	centrée réduite		0	1
2	beta 1	Bêta	-0,959	1er paramètre de forme=3	2e paramètre de forme=0,9	1	0,75

Au-delà de l'objet proprement dit de notre étude, nous savons que le nombre d'items joue un rôle important quant à la qualité des estimations des paramètres relatifs aux sujets (Hulin, Lissak & Drasgow, 1982). Il convenait donc d'introduire ce facteur dans l'étude pour saisir ses interactions avec les distributions envisagées. Trois modalités ont été fixées pour le nombre d'items : 10, 20 et 40.

Générateurs de données

Les matrices de réponses aux items (données alternatives : 0 pour une réponse incorrecte et 1 pour une réponse correcte) sont générées selon le modèle logistique à deux paramètres. Dans le modèle de Birnbaum, la probabilité qu'un sujet *j* donne une réponse correcte à un item *i* est fonction de la compétence (θ_j) du sujet, de la difficulté de l'item (b_i) et de la discrimination de l'item (a_i) (Hambleton, Swaminathan & Rogers, 1991).

$$P(x_{ij} = 1/\theta_j) = \frac{e^{[a_i(\theta_j - b_i)]}}{1 + e^{[a_i(\theta_j - b_i)]}}$$

La réponse x_{ij} d'un sujet *j* à un item *i* peut être simulée en calculant, dans un premier temps, par l'intermédiaire de l'équation (1), la probabilité de fournir une réponse correcte et en comparant, dans un deuxième temps, cette probabilité à un nombre aléatoire u_{ij} issu d'une distribution uniforme dans l'intervalle [0 ; 1] : si la probabilité de répondre correctement du sujet est supérieure à ou égale au nombre aléatoire u_{ij}, la réponse est codée comme correcte (1) ; par contre, si la probabilité de répondre correctement est inférieure au nombre aléatoire u_{ij}, la réponse est codée comme incorrecte (0) (Hulin *et al.,* 1982).

	Item1 a_1, b_1	*Item 2* a_2, b_2	... ...	*Item l* a_l, b_l
θ_1	$P_1(\theta_1) \sim u_{11}$	$P_2(\theta_1) \sim u_{12}$	...	$P_l(\theta_1) \sim u_{1l}$
θ_2	$P_1(\theta_2) \sim u_{21}$	$P_2(\theta_2) \sim u_{22}$	...	$P_l(\theta_2) \sim u_{2l}$
...	...	...	$P_i(\theta_j) \sim u_{ij}$	...
θ_n	$P_1(\theta_n) \sim u_{n1}$	$P_2(\theta_n) \sim u_{n2}$	...	$P_l(\theta_n) \sim u_{nl}$

Figure 1. ***Génération des matrices de réponses en fonction de la compétence des sujets (θ_j) et des paramètres des items (a_i, b_i)***

Les indices de difficulté des items ($b_1, ..., b_l$) sont des nombres aléatoires issus des distributions uniformes d'intervalle [-3 ;3] et [-3 ;-1] tandis que les indices de discrimination des items sont des nombres aléatoires issus des distributions uniformes d'intervalle [0,4 ;2,0]. Les valeurs de compétence des sujets ($\theta_1, ..., \theta_n$) sont, quant à elles. choisies de manière aléatoire et simple parmi les distributions normales et Bêta définies pour les besoins de l'étude. Ainsi, des échantillons de 4 000 sujets ont été générés.

Plan de simulation

Le croisement des deux variables prises en compte dans l'étude, c'est-à-dire le type de distribution de la compétence des sujets (deux modalités) et le degré d'homogénéité de la distribution des paramètres de difficulté des items (deux modalités), détermine quatre cas d'étude (tableau 2).

Tableau 2

Cas d'étude en fonction des distributions normales (N) et Bêta (B) de la compétence des sujets (θ_j) et des distributions uniformes (U) de la difficulté des items (b_i)

Cas d'étude	*Distribution des θ_j*	*Distribution des b_i*
CAS 1	N(0 ;1)	U[-3 ;3]
CAS 2	N(0 ;1)	U[-3 ;-1]
CAS 3	B(3 ;0,9)	U[-3 ;3]
CAS 4	B(3 ;0,9)	U[-3 ;-1]

Pour chacun de ces cas, dix répétitions ont été réalisées pour les trois modalités prises par la variable nombre d'items, soit un total de 120 matrices de réponses simulées.

Estimation des paramètres et analyse des résultats

Chacun des 120 échantillons simulés a été ajusté au modèle Birnbaum par la méthode du maximum de vraisemblance marginale pour l'estimation des paramètres des items. Ceux-ci étant fixés, la compétence des sujets a été déterminée par la méthode du maximum de vraisemblance classique.

À l'instar de l'étude de Hulin *et al.,* (1982) ou encore de celle d'Hambleton et Cook (1983), des coefficients de corrélation ont été utilisés pour quantifier et synthétiser la précision des estimations pour chaque échantillon, étant donné le caractère arbitraire des échelles. Nous avons donc analysé les coefficients de corrélation entre, d'une part, les valeurs théoriques fixées *a priori* lors de la génération des matrices de réponses et, d'autre part, les valeurs estimées *a posteriori* par la méthode du maximum de vraisemblance marginale.

En vue de la réalisation des analyses de la variance, la vérification des conditions de normalité et d'homoscédasticité effectuée sur les résidus réduits des données initiales a imposé l'application d'une transformation de variable pour réduire les écarts à la normale et l'hétérogénéité des variances. Étant donné la nature des observations réalisées, nous avons utilisé la transformation argument tangente hyperbolique, particulièrement adaptée aux problèmes de corrélation (Dagnelie, 1998) et définie comme suit :

$$z = \frac{1}{2}\ln\frac{1+R}{1-R}$$

Des analyses de la variance à deux critères de classification (modèle croisé fixe) ont été ensuite réalisées sur les coefficients de corrélation transformés de chaque échantillon. En complément de l'analyse de la variance, nous avons procédé à des comparaisons *post hoc* des moyennes (test HSD de Tukey) pour affiner l'étude. Les analyses ont donc porté sur les variables transformées en fonction du carré moyen résiduel des modèles de l'analyse de la variance qui ont précédé la phase de comparaison au seuil de signification 0,05. Les résultats sont cependant présentés en fonction des données initiales après avoir appliqué la transformation inverse.

L'analyse des résultats comporte trois parties qui traiteront successivement de la précision des estimations de la discrimination et de la difficulté des items ainsi que de la précision des estimations de la compétence des sujets. Dans chacune de ces parties, nous avons donc procédé de manière systématique à un examen des données initiales, à la réalisation d'une analyse de variance après avoir vérifié les conditions d'application et, le cas échéant, à une comparaison des moyennes relatives aux facteurs étudiés.

Résultats

Paramètres de discrimination des items

Les moyennes des coefficients de corrélation entre les valeurs théoriques et estimées de la discrimination des items (tableau 3) sont très élevées, quel que soit l'échantillon considéré. Un minimum est observé lorsque le trait latent dans la population est distribué selon la distribution Bêta de faible dispersion, les indices de difficulté des items sont relativement homogènes (cas 3) et le nombre d'items est faible (dix items).

Tableau 3
Moyennes et écarts types des coefficients de corrélation relatifs aux estimations de la discrimination des items

CAS	*NBITEM*	*Moyennes*	*Écarts types*	*N*
1	10	0,976	0,019	10
	20	0,985	0,010	10
	40	0,985	0,006	10
	Total	0,982	0,013	30
2	10	0,980	0,008	10
	20	0,978	0,010	10
	40	0,986	0,007	10
	Total	0,981	0,009	30
3	10	0,875	0,051	10
	20	0,911	0,051	10
	40	0,922	0,018	10
	Total	0,903	0,046	30
4	10	0,910	0,051	10
	20	0,940	0,020	10
	40	0,959	0,013	10
	Total	0,936	0,037	30
Total	10	0,935	0,058	40
	20	0,954	0,041	40
	40	0,963	0,029	40
	Total	0,951	0,045	120

De prime abord, les écarts observés par rapport à la situation de référence (cas 1) sont très faibles pour le deuxième cas mais sont assez marqués pour les troisième et quatrième cas. Si l'on considère à présent les moyennes relatives aux différentes modalités de la variable «nombre d'items» à l'intérieur de chaque cas, celles-ci sont quasiment identiques pour les deux premiers cas alors qu'elles semblent augmenter en fonction du nombre d'items pour les deux derniers.

L'interaction entre les deux facteurs principaux issus du modèle d'analyse de la variance croisé fixe n'étant pas significative, les conclusions que l'on peut en tirer sont claires. L'effet conjugué de la distribution du trait latent dans la population et du degré d'homogénéité des indices de difficulté des items sur les coefficients de corrélation entre les valeurs théoriques et estimées de la discrimination des items ainsi que l'effet du nombre d'items sont significatifs (tableau 4).

Tableau 4

Étude de l'effet des cas envisagés et du nombre d'items sur les estimations de la discrimination des items (analyse de la variance)

Source	*Somme des carrés*	*Degrés de liberté*	*Carré moyen*	*F*	*p*
Modèle complet	520,339	12	43,362	618,296	0,000
CAS	18,563	3	6,188	88,230	0,000
NBITEM	0,915	2	0,457	6,522	0,002
CAS * NBITEM	0,405	6	0,067	0,962	0,455
Erreur	7,574	108	0,070		
Total	527,914	120			

Après comparaison des moyennes des coefficients de corrélation (tableau 5), il apparaît que les cas 1 et 2, c'est-à-dire les cas où la distribution du trait latent dans la population est normale, donnent lieu aux estimations les plus précises pour les paramètres de discrimination des items. Par contre, avec un coefficient de corrélation de 0,912, le cas 3, où la distribution du trait latent dans la population n'est pas distribuée normalement et où les paramètres de difficulté des items sont relativement hétérogènes, génère les estimations les moins précises.

Tableau 5
Comparaison des moyennes relatives aux cas envisagés (HSD Tukey)

		Sous-ensembles		
CAS	*N*	*1*	*2*	*3*
3	30	0,912		
4	30		0,945	
2	30			0,983
1	30			0,985

En ce qui concerne l'influence du nombre d'items sur la précision des estimations des paramètres de discrimination des items (tableau 6), les résultats sont contradictoires dans la mesure où apparaît un phénomène de chevauchement. En effet, si les moyennes estimées au départ de 10 et 40 items semblent significativement différentes, elles ne diffèrent pas d'une même troisième calculée au départ de 20 items.

Tableau 6
Comparaison des moyennes relatives aux nombres d'items (HSD Tukey)

		Sous-ensembles	
NBITEM	*N*	*1*	*2*
10	40	0,959	
20	40	0,967	0,967
40	40		0,973

Paramètres de difficulté des items

Contrairement aux résultats obtenus pour les paramètres de discrimination des items, la moyenne la plus faible des coefficients de corrélation est observée pour le cas 4, caractérisé par une distribution Bêta du trait latent dans la population et par une faible dispersion des paramètres de difficulté des items (tableau 7). Le maximum est, quant à lui, toujours atteint pour le cas 1 qui a été choisi comme témoin pour la comparaison.

Tableau 7
Moyennes et écarts types des coefficients de corrélation relatifs aux estimations de la difficulté des items

CAS	*NBITEM*	*Moyennes*	*Écarts types*	*N*
1	10	0,998	0,001	10
	20	0,998	0,001	10
	40	0,999	0,000	10
	Total	0,998	0,001	30
2	10	0,975	0,035	10
	20	0,975	0,015	10
	40	0,983	0,007	10
	Total	0,978	0,022	30
3	10	0,992	0,004	10
	20	0,992	0,004	10
	40	0,993	0,001	10
	Total	0,993	0,003	30
4	10	0,810	0,119	10
	20	0,897	0,044	10
	40	0,900	0,015	10
	Total	0,869	0,083	30
Total	10	0,944	0,099	40
	20	0,966	0,047	40
	40	0,969	0,041	40
	Total	0,959	0,068	120

Les écarts par rapport à la situation de référence se marquent essentiellement pour le quatrième cas et, dans une moindre mesure, pour le deuxième cas. Ce n'est que pour le dernier cas qu'*a priori* les moyennes augmentent en fonction du nombre d'items employés pour l'ajustement des modèles.

Ici aussi, l'interaction entre les facteurs «cas envisagé» et «nombre d'items» n'est pas significative (tableau 8). Cette dernière constatation permet d'interpréter aisément les effets des facteurs principaux, ce qui n'est pas significatif pour le nombre d'items mais bien en ce qui concerne les cas envisagés dans l'étude.

Tableau 8
Étude de l'effet des cas envisagés et du nombre d'items sur les estimations de la difficulté des items (analyse de la variance)

Source	*Somme des carrés*	*Degrés de liberté*	*Carré moyen*	*F*	*p*
Modèle complet	865,281	12	72,107	869,658	0,000
CAS	78,392	3	26,131	315,154	0,000
NBITEM	0,130	2	0,065	0,782	0,460
CAS * NBITEM	0,981	6	0,164	1,973	0,076
Erreur	8,955	108	0,083		
Total	874,236	120			

Ainsi, il apparaît clairement que le quatrième cas, le moins favorable, conduit à des estimations de la difficulté des items nettement moins précises que les autres tandis que les cas 1 et 3, caractérisés par une distribution initialement hétérogène des paramètres de difficulté des items, conduisent aux estimations les plus précises (tableau 9).

Tableau 9
Comparaison des moyennes relatives aux cas envisagés (HSD Tukey)

		Sous-ensembles			
CAS	*N*	*1*	*2*	*3*	*4*
4	30	0,883			
2	30		0,983		
3	30			0,993	
1	30				0,999

Estimations de la compétence des sujets

La fluctuation de la qualité des estimations des paramètres relatifs aux items en fonction des variables indépendantes envisagées dans cette étude a des conséquences importantes sur la précision des estimations de la compétence des sujets puisque ces dernières sont réalisées à partir des premières. Ainsi, toute imprécision générée lors de la première phase d'ajustement des modèles aux données peut se répercuter lors de l'estimation de la compétence des sujets.

Le premier cas, employé comme témoin, conduit naturellement aux coefficients de corrélation les plus élevés, avec une moyenne de 0,87 (tableau 10). Pour le quatrième cas, le plus défavorable puisque le trait latent y est réparti selon une distribution Bêta asymétrique et que la distribution de la difficulté des items présente une faible dispersion, la moyenne des coefficients de corrélation ne dépasse pas 0,70, ce qui dénote une qualité médiocre des estimations de la compétence des sujets. Les deuxième et troisième cas semblent, quant à eux, conduire à des qualités d'estimation quasi similaires avec des coefficients de corrélation moyens proches de 0,80.

Tableau 10
Moyennes et écarts types des coefficients de corrélation relatifs aux estimations de la compétence des sujets

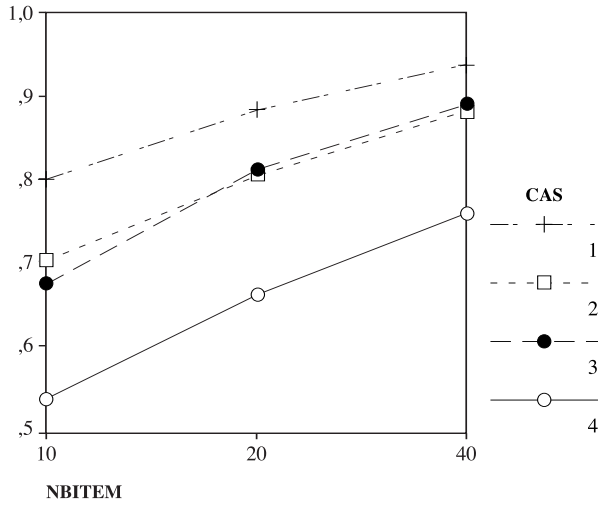
CAS	*NBITEM*	*Moyennes*	*Écarts types*	*N*
1	10	0,797	0,037	10
	20	0,884	0,018	10
	40	0,935	0,005	10
	Total	0,872	0,062	30
2	10	0,704	0,033	10
	20	0,804	0,018	10
	40	0,880	0,005	10
	Total	0,796	0,076	30
3	10	0,677	0,040	10
	20	0,812	0,027	10
	40	0,889	0,013	10
	Total	0,793	0,093	30
4	10	0,541	0,039	10
	20	0,663	0,021	10
	40	0,760	0,020	10
	Total	0,655	0,095	30
Total	10	0,680	0,100	40
	20	0,791	0,084	40
	40	0,866	0,066	40
	Total	0,779	0,114	120

À l'exception du quatrième cas, pour lequel les coefficients de corrélation sont faibles quel que soit le nombre d'items utilisés, au-delà de 20 items la valeur des coefficients de corrélation entre les valeurs théoriques et estimées de la compétence des sujets dépasse 0,80 et approche 0,90 lorsque 40 items

sont employés pour l'estimation. Le maximum est enregistré bien évidemment pour la situation de référence lorsque l'ajustement du modèle est effectué avec 40 items. Le minimum apparaît lorsque la distribution du trait latent dans la population n'est pas normale, que la distribution de la difficulté des items est relativement homogène et que les estimations sont réalisées à partir de dix items, donnant lieu à un coefficient de corrélation de 0,54.

Un premier modèle d'analyse de la variance à deux critères de classification croisés fixes a été réalisé et montre clairement que le degré de signification de l'interaction entre les deux facteurs principaux enlève toute valeur aux tests réalisés sur les facteurs principaux. Il sera donc plus prudent de tester la signification de ces facteurs séparément pour leurs différentes modalités.

L'étude détaillée de l'interaction (graphique 2) montre que si le cas 2 donne lieu à des estimations plus précises de la compétence des sujets lorsque l'estimation est réalisée avec dix items en comparaison au cas 3, la tendance s'inverse pour 20 et 40 items. Cependant, les différences étant minimes, on ne peut conclure qu'à la proximité de ces deux cas. On peut également noter que les disparités entre les cas d'étude sont plus faibles lorsque le nombre d'items augmente.



Graphique 2. ***Moyennes marginales des coefficients de corrélation en fonction des cas envisagés et du nombre d'items***

Le deuxième modèle d'analyse de la variance où les effets du facteur «cas» ont été testés pour chacune des modalités de la variable «nombre d'items» (tableau 11) atteste, quel que soit le nombre d'items utilisés, qu'il existe des différences significatives quant aux moyennes des coefficients de corrélation entre les valeurs théoriques et estimées de la compétence des sujets. De même, un troisième modèle d'analyse de la variance où les effets du facteur «nombre d'items» ont été testés pour chacune des modalités de la variable «cas» atteste également, quel que soit le cas envisagé, l'existence de différences significatives pour ces mêmes paramètres.

Tableau 11
Étude de l'effet des cas envisagés et du nombre d'items sur les estimations de la compétence des sujets (deuxième tableau de l'analyse de la variance)

Source	*Somme des carrés*	*Degrés de liberté*	*Carré moyen*	*F*	*p*
NBITEM	5,420	2	2,710	637,983	0,000
CAS / 10 items	1,219	3	0,406	95,665	0,000
CAS / 20 items	1,815	3	0,605	142,411	0,000
CAS / 40 items	2,494	3	0,831	195,761	0,000
Erreur	0,459	108	0,004		
CAS	5,381	3	1,794	422,326	0,000
NBITEM / cas 1	1,812	2	0,906	213,320	0,000
NBITEM / cas 2	1,241	2	0,621	146,138	0,000
NBITEM / cas 3	1,747	2	0,873	205,593	0,000
NBITEM / cas 4	0,766	2	0,383	90,199	0,000
Erreur	0,459	108	0,004		

Après comparaison des moyennes (tableaux 12 et 13), on peut constater qu'en dehors du cas le plus favorable où le trait latent dans la population est distribué normalement et où la distribution de la difficulté des items est relativement hétérogène, les estimations de la compétence des sujets sont encore relativement précises si l'une ou l'autre de ces conditions ne sont pas remplies. Par contre, lorsque la distribution du trait latent n'est plus normale et que, simultanément, la distribution de la difficulté des items présente une faible dispersion, les estimations ne sont pas suffisamment précises puisque la moyenne des coefficients de corrélation entre les valeurs théoriques et estimées ne dépasse pas 0,70.

Tableau 12
Comparaison des moyennes relatives aux cas envisagés (HSD Tukey)

		Sous-ensembles		
CAS	*N*	*1*	*2*	*3*
4	30	0,665		
2	30		0,808	
3	30		0,810	
1	30			0,885

Conformément à l'examen des données initiales, si l'on considère le nombre d'items, les moyennes des coefficients de corrélation relatives aux trois modalités envisagées se distinguent nettement les unes des autres.

Tableau 13
Comparaison des moyennes relatives aux nombres d'items (HSD Tukey)

		Sous-ensembles		
NBITEM	*N*	*1*	*2*	*3*
10	40	0,693		
20	40		0,805	
40	40			0,879

Conclusion

Bien que les choix des modalités des paramètres qui sont envisagés dans cette étude limitent la portée des conclusions que l'on peut en tirer, un certain nombre de constats peuvent être établis et quantifiés.

Si ce n'est lorsque la distribution du trait latent dans la population n'est pas normale, les estimations des paramètres de discrimination des items dans le modèle de Birnbaum sont relativement précises. Et bien que le nombre d'items ait un effet significatif sur la qualité de ces estimations, l'ampleur des différences observées est peu importante.

Dans le cas le moins favorable, où le trait latent dans la population n'est pas distribué normalement et où la distribution de la difficulté des items est relativement homogène, la qualité d'estimation des paramètres de difficulté des items est nettement inférieure aux autres cas. Par contre, le nombre d'items utilisés pour l'ajustement du modèle n'a pas d'effet significatif sur la précision des estimations de ces derniers paramètres.

En ce qui concerne l'estimation de la compétence des sujets, le nombre d'items, le type de distribution du trait latent dans la population et le degré d'homogénéité de la distribution de la difficulté des items ont un effet très important sur la qualité d'estimation. Ainsi, lorsque le trait latent dans la population n'est pas distribué normalement et que l'homogénéité de la distribution de la difficulté des items est relativement importante, la précision des estimations est faible. De même, lorsque le nombre d'items utilisés pour ajuster le modèle de Birnbaum est peu élevé, la moyenne des coefficients de corrélation entre les valeurs théoriques et estimées de la compétence des sujets s'en trouve diminuée.

Dès lors, dans quelle mesure peut-on avoir confiance envers l'estimation de la compétence des sujets par le modèle de Birnbaum ?

Au regard des résultats obtenus dans cette étude, nous sommes en droit de nous interroger sur l'exactitude des estimations obtenues et, *a fortiori*, sur la pertinence des décisions qui en découlent. Ainsi, les simulations réalisées montrent notamment que, à l'instar des conditions d'application que l'on trouve dans certaines évaluations sur le plan national (items peu nombreux et dont les degrés de difficulté sont relativement homogènes), la précision des estimations réalisées est trop faible pour garantir un traitement équitable des performances des élèves. De manière moins dramatique, si l'on considère que les compétences des élèves sont distribuées normalement dans la population évaluée, la qualité des estimations se révèle quand même insuffisante pour des évaluations dont les enjeux sont importants.

Tableau 14
Distribution des écarts entre les valeurs théoriques et estimées de la compétence des sujets dans les cas où le trait latent dans la population est distribué normalement mais que la distribution de la difficulté des items est relativement homogène et que les estimations sont réalisées à partir de dix items

Classe	*Nombre de sujets*	*Pourcentage*
]-2,5 ; -2,0]	3	0,1 %
]-2,0 ; -1,5]	30	0,8 %
]-1,5 ; -1,0]	204	5,1 %
]-1,0 ; -0,5]	630	15,8 %
]-0,5 ; 0,0]	1174	29,4 %
]0,0 ; 0,5]	1101	27,5 %
]0,5 ; 1,0]	602	15,1 %
]1,0 ; 1,5]	187	4,7 %
]1,5 ; 2,0]	52	1,3 %
]2,0 ; 2,5]	13	0,3 %
]2,5 ; 3,0]	1	0,0 %
]3,0 ; 3,5]	2	0,1 %
]3,5 ; 4,0]	1	0,0 %
Total	4000	100 %

Concrètement, si l'on considère les écarts entre les valeurs réelles de la compétence des élèves et celles estimées par le modèle de Birnbaum (tableau 14), on peut constater que, par exemple, dans le cas où le trait latent dans la population est distribué normalement mais que la distribution de la difficulté des items est relativement homogène et que les estimations sont réalisées à partir de dix items, près de 43 % des valeurs estimées de la compétence des élèves sont entachées d'une erreur supérieure à 0,5 (sur une échelle normale standard) en valeur absolue. Ainsi, pour un nombre important d'élèves, les conclusions tirées au départ de l'estimation de leur compétence peuvent se révéler totalement inexactes. L'ampleur des erreurs liées aux estimations peut avoir, dans ce cas, des conséquences graves puisqu'un élève dont la compétence réelle est au départ satisfaisante pourra être considéré comme médiocre si l'erreur qui affecte son score est suffisamment importante.

Par contre, si la distribution de la difficulté des items est relativement hétérogène et si on utilise un grand nombre d'items pour évaluer la compétence des sujets (tableau 15), le pourcentage d'élèves dont l'estimation de la compétence est entachée d'une erreur supérieure à 0,5 est inférieure à 14%. Et aucun écart entre les valeurs réelles de la compétence des sujets et les valeurs estimées n'excède 1,5. Autrement dit, les risques de tirer des conclusions erronées sur les capacités des élèves sont nettement réduits mais ne sont pas nuls.

Tableau 15
Distribution des écarts entre les valeurs théoriques et estimées de la compétence des sujets dans les cas où le trait latent dans la population est distribué normalement, que la distribution de la difficulté des items est relativement hétérogène et que les estimations sont réalisées à partir de 40 items

Classe	*Nombre de sujets*	*Pourcentage*
]-2,5 ; -2,0]	0	0,0%
]-2,0 ; -1,5]	0	0,0%
]-1,5 ; -1,0]	7	0,2%
]-1,0 ; -0,5]	282	7,1%
]-0,5 ; 0,0]	1773	44,3%
]0,0 ; 0,5]	1680	42,0%
]0,5 ; 1,0]	251	6,3%
]1,0 ; 1,5]	7	0,2%
]1,5 ; 2,0]	0	0,0%
]2,0 ; 2,5]	0	0,0%
]2,5 ; 3,0]	0	0,0%
]3,0 ; 3,5]	0	0,0%
]3,5 ; 4,0]	0	0,0%
Total	4000	100%

Dans le cadre de procédures d'orientation ou de certification, le degré de précision des estimations réalisées au moyen des MRI revêt un caractère particulièrement important puisque, dans ces cas, rappelons que les résultats des élèves peuvent avoir des conséquences graves sur leur avenir. Aussi, si l'on se place dans une perspective de justice scolaire, il est nécessaire de garantir à chaque enfant un traitement équitable de ses performances en assurant, dans un premier temps, une qualité optimale des mesures réalisées lors de la conception et du traitement des épreuves.

Les résultats de cette étude montrent clairement que si nous ne pouvons agir sur la nature de la distribution du trait latent au sein d'une population testée, il est possible de réduire les imprécisions des estimations en jouant sur le nombre d'items utilisés et en veillant à ce que les paramètres de difficulté des items choisis pour l'évaluation soient suffisamment variés pour couvrir un domaine large de compétence.

Par ailleurs, on peut constater aussi bien théoriquement (Dagnelie, 1998) qu'empiriquement (Hambleton & Rovinelli, 1973) que lorsque l'ajustement se réalise à partir d'un échantillon de sujets dont l'effectif est très élevé, la puissance des tests d'ajustement est telle que les plus infimes écarts par rapport aux modèles théoriques sont détectés, impliquant de la sorte le rejet de nombreux items. La situation devient particulièrement préoccupante lorsque le nombre d'items restant est faible puisque, dans ce cas, la qualité de l'estimation de la compétence des sujets s'en trouve nettement diminuée.

Aussi, ne serait-il pas plus judicieux, dans de telles situations, de conserver dans l'ajustement du modèle final des items qui, bien que ne s'ajustant pas parfaitement au modèle, présentent des écarts infimes par rapport au modèle théorique. L'impact de la suppression de tels items devra faire l'objet d'études plus approfondies pour déterminer l'optimum de la précision des estimations de la compétence des sujets en fonction du rejet ou du maintient d'items mal ajustés et du nombre total d'items qui ont servi à l'ajustement du modèle.

RÉFÉRENCES

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & M. Novick (dir.), *Statistical Theories of Mental Test Scores* (pp. 397-549). Reading, MA : Addison-Wesley.

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

Dagnelie, P. (1998). *Statistique théorique et appliquée* (tome 2). Bruxelles : De Boeck.

Épreuves standardisées 2000/2001 (2001). *Documentation des résultats des épreuves standardisées*. Luxembourg : Ministère de l'Éducation nationale, de la Formation professionnelle et des Sports.

Follmann, D. (1988). Consistent estimation in the Rasch model based on nonparametric margins. *Psychometrika, 53*(4), 553-562.

Habermann, S. (1977). Maximum likelihood estimates in exponential response models. *The Annals of Statistics, 5*, 815-841.

Hambleton, R.K. (1994). Item response theory : A broad psychometric framework for measurement advances. *Psicothema, 6*, 535-556.

Hambleton, R.K., & Cook, L.L. (1983). Robustness of item response models and effects of test length and sample size on the precision of ability estimates. In D. Weiss (éd.), *New horizons in testing* (pp. 31-49). New York : Academic Press.

Hambleton, R.K., & Rovinelli, R.J. (1973). A Fortran IV program for generating examinee response data from logistic test models. *Behavioral Science, 17*, 73-74.

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

Hulin, C.L., Lissak, R.I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves : A Monte Carlo study. *Applied Psychological Measurement, 6*, 249-260.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. (expanded edition). Chicago : The University of Chicago Press. (Œuvre originale publiée en 1960.)

Vrignaud, P. & Chartier, P. (1999). *Quand les modèles de mesure deviennent réducteurs : apports et limites des modèles de réponse à l'item pour les comparaisons internationales.* Paris : Service de Recherche de l'INETOP.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1996). *BILOG-MG*. Chicago, IL : Scientific Software International.