La multidimensionnalité dans l'évaluation PISA 2003 : la place d'une dimension générale

Elodie Pools – elodie.pools@uliege.be
Assistante – Université de Liège

Christian Monseur – cmonseur@uliege.be
Professeur – Université de Liège

Pour citer cet article : Pools, E., & Monseur, C. (2018). La multidimensionnalité dans l'évaluation PISA 2003 : la place d'une dimension générale. *Évaluer. Journal international de recherche en éducation et formation, 4*(3), 21-45.

Résumé

Le Programme International de Suivi des Acquis des élèves (PISA) évalue les compétences en littéracie des élèves de 15 ans principalement dans trois domaines (lecture, sciences et mathématiques) par le biais d'un test cognitif. Chacun de ses items est supposé ne cibler que la compétence du domaine qu'il est sensé mesurer (hypothèse de multidimensionnalité *between-item*). En d'autres termes, tous les items ne contribuent qu'à mesurer une seule compétence. Or, les performances des élèves dans les trois domaines évalués corrèlent fortement entre elles, traduisant un recouvrement important entre les différentes dimensions, ce qui questionne cette hypothèse de multidimensionnalité *between-item*.

Cet article présente trois structures latentes sous-jacentes à l'évaluation PISA 2003: un modèle unidimensionnel, un modèle multidimensionnel *between-item* et un modèle multidimensionnel *within-item* où chaque item mesure à la fois une dimension spécifique et une dimension générale. Ces modèles sont testés par le biais de modèles de réponse à l'item dans 17 pays.

Les résultats indiquent que les modèles multidimensionnels s'ajustent mieux aux données que le modèle unidimensionnel, avec, dans la majorité des pays, un léger avantage pour le modèle *within-item*. Ce modèle *within-item* révèle qu'en moyenne, les items ont une meilleure discrimination de la dimension générale que des dimensions spécifiques ; ces compétences spécifiques doivent cependant être inclues dans le modèle, le modèle unidimensionnel ayant le moins bon ajustement statistique aux données.

Cette structure complexe pourrait avoir des conséquences notamment sur les associations entre les estimations de la performance et les variables contextuelles telles que décrites dans les rapports de l'OCDE.

Mots-clés

Théorie de réponse à l'item ; évaluation à large échelle ; multidimensionnalité.

Abstract

The Programme for International Student Assessment (PISA) assesses 15 years old students' ability in litteracy mainly in three domains (reading, mathematics and science). Students are assessed by means of a cognitive test in which items are supposed to assess only the domain they belong to (between-item multidimensionality assumption). Each item is thus supposed to measure only one dimension. These students performance estimates highly correlate and these strong correlations might therefore reflect an overlapping of the assessed dimensions, which rises questions on the between-item multidimensionality assumption.

This article presents the results of three latent structures to PISA 2003 cognitive data: a unidimensional model, a model with between-item multidimensionality and a model with within-item multidimensionality. In this last model, each item measures an ability specific to the domain assessed by the item and a general ability that is measured by all the items of the cognitive test. Item Response Theory models are used to fit these models on the data of 17 countries.

Results show that the multidimensional models better fit the data than the unidimensional model, the best-fitting model being the within-item model in most of the countries. Item discrimination parameters of the within-item model are, on average, higher on the general ability than on the specific abilities; these specific abilities should nevertheless be included in the model as the unidimensional model has the worst fit.

This within-item model may have implications concerning, among other things, the association between ability estimates and background variables as reported in the OECD's reports.

Keywords

Item Response Theory; large scale assessment; multidimensionality.

1. Introduction

Le Programme International pour le Suivi des Acquis des élèves (PISA) est une évaluation des compétences des élèves de 15 ans menée tous les 3 ans par l'Organisation de Coopération et de Développement Economiques (OCDE). À travers cette enquête internationale, l'OCDE ne vise pas à déterminer le niveau de compétence des élèves sur des points précis du curriculum (qui, de fait, varie d'un pays à l'autre) mais plutôt à mesurer le degré de littéracie principalement dans trois domaines (les mathématiques, la lecture et les sciences) ainsi que dans des domaines spécifiques à certains cycles, comme la résolution de problèmes en 2003. PISA définit la littéracie comme la capacité des étudiants (i) à appliquer des connaissances et savoirs provenant de domaines clés et (ii) à analyser, raisonner et communiquer de manière effective en résolvant des problèmes dans une diversité de situations (OCDE, 2016, p.11). Les estimations de ces compétences sont ensuite utilisées pour dériver des indicateurs d'efficacité et d'équité des systèmes éducatifs.

L'évaluation cognitive des élèves est au cœur de PISA. Les étudiants échantillonnés sont invités à répondre à des items tels ceux présentés à la figure 1. Ces items, extraits de PISA 2003 (OCDE, 2004), sont fortement contextualisés. Prenons l'exemple de la question 11. Au niveau du contenu mathématique, elle présente directement l'ensemble des informations requises à la résolution du problème et seules les opérations arithmétiques de base sont sollicitées (multiplication et division). Si cet item est relativement simple sur le plan mathématique, son haut degré de contextualisation le rend complexe. Les étudiants doivent formuler un modèle mathématique correct au départ de ces données contextualisées et utiliser leurs compétences de raisonnement, d'argumentation, de résolution de problèmes ainsi que de communication afin de produire et justifier leur réponse (OCDE, 2004). Ainsi, au-delà de compétences propres aux mathématiques, cet item mesure également d'autres compétences, telle la capacité à comprendre une situation contextualisée.

TAUX DE CHANGE

Mademoiselle Mei-Ling, de Singapour, prépare un séjour de 3 mois en Afrique du Sud dans le cadre d'un échange d'étudiants. Elle doit changer des dollars de Singapour (SGD) en rands sud-africains (ZAR).

Question 9

Mei-Ling a appris que le taux de change entre le dollar de Singapour et le rand sud-africain est de: 1 SGD = 4,2 ZAR. Mei-Ling a changé 3 000 dollars de Singapour en rands sud-africains à ce taux de change.
Combien Mei-Ling a-t-elle reçu de rands sud-africains ?

Question 10

Lorsque Mei-Ling rentre à Singapour après 3 mois, il lui reste 3 900 ZAR. Elle les reconvertit en dollars de Singapour, constatant que le taux de change a évolué et est à présent de : 1 SGD = 4,0 ZAR.
Combien Mei-Ling reçoit-elle de dollars de Singapour ?

Question 11

Durant ces 3 mois, le taux de change a changé de 4.2 à 4.0 ZAR par SGD.
Était-ce en faveur de Mei-Ling que le taux de change soit désormais de 4.0 ZAR au lieu de 4.2 ZAR alors qu'elle échange ses rands sud-africains en dollars de Singapour ? Donne une explication pour justifier ta réponse.

Figure 1. Exemple d'unité conçue pour l'échelle PISA de culture mathématique « Quantité » : Taux de change. Traduit de "Learning for tomorrow's world: first results from PISA 2003", p. 75, OCDE, 2004, Paris, OECD Publishing

Les réponses des élèves aux différents items sont ensuite utilisées afin d'estimer leurs compétences. Ce faisant, l'OCDE fait l'hypothèse que les items ne mesurent que la compétence du domaine dont ils relèvent. En d'autres termes, la probabilité de réussir un item du questionnaire tel que l'item 11 sur les taux de change n'est fonction que de la compétence spécifique au domaine auquel appartient l'item (ici, les mathématiques), aucune compétence commune à l'ensemble des items n'influençant cette probabilité. Les items de mathématiques ne mesurent donc que la compétence en mathématiques, ceux de lecture que la compétence en lecture, et ainsi de suite. Les compétences ainsi estimées dans chacun des domaines sont fortement corrélées : dans les pays de l'OCDE, en moyenne, les corrélations latentes sont comprises entre 0.77 (mathématiques et lecture) et 0.89 (mathématiques et résolution de problèmes) (OCDE, 2005). Ces corrélations élevées peuvent traduire un recouvrement entre les compétences (Blömeke, Houan & Suhl, 2014), ce recouvrement pouvant s'expliquer par la présence d'une dimension générale mesurée par tous les items.

La présence d'une dimension générale, communément évaluée par l'ensemble des items sans distinction de domaine, peut donc se justifier tant par le cadre conceptuel de l'évaluation que par les possibles recouvrements entre les compétences estimées. Par ailleurs, on ne peut pas exclure que la performance observée par l'intermédiaire des réponses ait pu être influencée par un engagement variable d'un élève à l'autre, d'autant plus que l'évaluation PISA ne comporte aucun enjeu pour les élèves évalués.

Le présent article propose une comparaison de trois structures latentes de la (ou des) compétence(s) évaluée(s) par le questionnaire cognitif de PISA 2003 ; ce dernier, outre les trois domaines susmentionnés, mesure également la compétence en résolution de problèmes. Le premier modèle est un modèle unidimensionnel (figure 2a) dans lequel tous les items ne mesurent que cette dimension générale, indépendamment des domaines auxquels ils appartiennent. Ce modèle fait l'hypothèse que les réponses fournies par les élèves relèvent d'une seule dimension générale. Dans le second modèle (figure 2b), la réponse à chaque item ne dépend que de la dimension propre à son domaine et aucun trait latent général n'est impliqué. Ce modèle, dénommé multidimensionnel *between-item*, est mis en œuvre depuis le premier cycle de PISA pour dériver les performances des élèves. Enfin, le troisième modèle (figure 2c) est un modèle multidimensionnel proposant une structure complexe dans laquelle les items de mathématiques, lecture et sciences dépendent à la fois de la compétence spécifique à leur domaine respectif mais aussi d'une dimension commune à tous les items. Afin que le modèle puisse être identifié, les items de résolution de problèmes ne peuvent que mesurer la dimension générale.

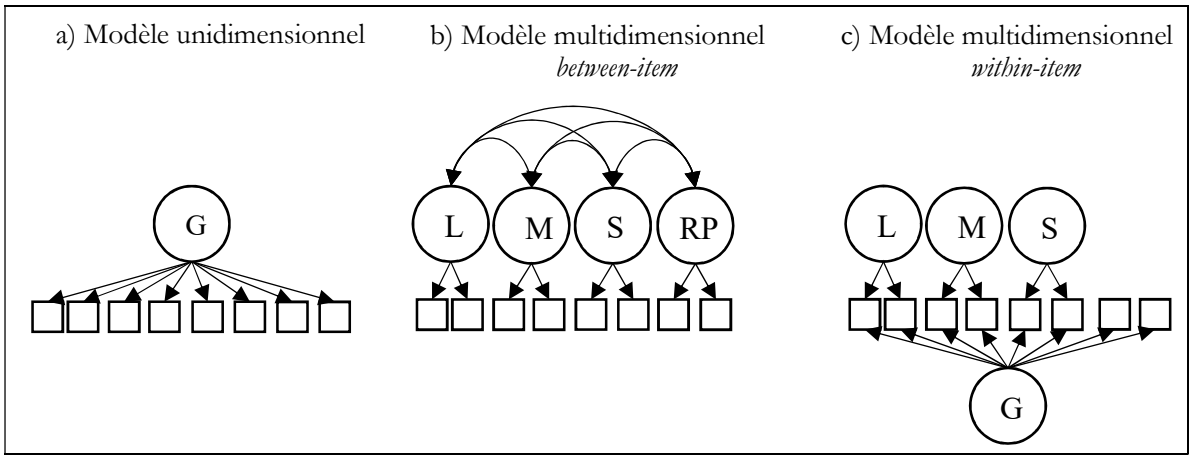

Figure 2. Modèles a) unidimensionnel, b) multidimensionnel *between-item* et c) multidimensionnel *within-item* (G représente la dimension générale, L la compétence dans le domaine de la lecture, M en mathématiques, S en sciences, RP en résolution de problèmes ; les carrés représentent les items et les cercles les traits latents)

1.1. Une analyse de la dimensionnalité du test

Mesurer une caractéristique d'un individu, telle sa compétence en mathématiques, c'est décrire un (et un seul) attribut de cette personne : la mesure est donc toujours unidimensionnelle (Wright & Masters, 1982). Mesurer une caractéristique d'un individu, c'est estimer sa position sur cette variable par le biais d'un instrument de mesure, à savoir un test composé d'items (Wright & Masters, 1982). Si le trait latent mesuré est unidimensionnel, l'outil de mesure ne l'est pas forcément (Wang, Wilson & Adams, 1997). Wang *et al.* (1997) définissent deux niveaux de multidimensionnalité de l'outil de mesure : le test et l'item. Ces deux niveaux permettent de définir trois types de tests : les tests unidimensionnels, les tests multidimensionnels *between-item* et enfin les tests multidimensionnels *within-item*.

Les tests unidimensionnels sont composés d'items unidimensionnels, tous les items contribuant à mesurer une seule et même dimension. D'autres tests peuvent être multidimensionnels, soit parce qu'ils sont constitués de différents sous-tests mesurant chacun une dimension distincte, soit parce qu'ils regroupent des items mesurant plusieurs dimensions. Les modèles multidimensionnels *between-item* et *within-item* postulent chacun un niveau de multidimensionnalité différent, ces différences traduisant différentes hypothèses quant au fonctionnement du test et amenant à différentes interprétations des échelles en émanant.

Lorsqu'un test est composé de différents sous-tests (chacun évaluant une dimension différente) au sein desquels chaque item ne mesure qu'une seule et même dimension, la multidimensionnalité se situe entre les items (*between-item* en anglais) et non au sein des items, ces derniers étant unidimensionnels (Wang *et al.*, 1997). Cette structure simple est donc recommandée quand chaque dimension affecte un ensemble différent d'items (Wu & Adams, 2006). Ainsi, pour l'item « Taux de change », dans le modèle *between-item* tel qu'il est implémenté dans PISA depuis le premier cycle (PISA2000), seule la compétence en mathématiques de l'étudiant influence sa réponse à l'item (Figure 2b). Un des avantages de ce modèle tient à la facilité de l'interprétation des traits latents : chaque item ne mesurant qu'une seule dimension, les différents traits latents peuvent être définis en fonction des items correspondants (Blömeke *et al.*, 2014; Hartig & Höhler, 2009). Enfin, lorsque les

dimensions ne sont pas orthogonales, le modèle *between-item* est plus efficace que des implémentations répétées de modèles unidimensionnels sur chaque sous-test pour estimer avec précision la compétence des individus et les corrélations latentes entre dimensions (Sheng & Wikle, 2007; Wang, Chen & Cheng, 2004 ; Wang *et al.*, 1997).

D'autres tests peuvent être constitués d'items multidimensionnels : ces items mesurent simultanément plusieurs dimensions et la multidimensionnalité se situe donc au sein des items (*within-item* en anglais) (Wang *et al.*, 1997). Le modèle multidimensionnel *within-item* postule une structure complexe des traits latents : plusieurs dimensions influencent la probabilité de réussir un même item, chacune de celles-ci étant requise à la réussite de la tâche (Hartig & Höhler, 2009). Dans l'exemple de l'item sur les taux de change, la réussite de cet item est désormais fonction de la compétence spécifique en mathématiques des étudiants mais aussi de leur niveau sur la dimension générale (Figure 2c). Si ce trait latent général a un impact sur la réussite de l'item, la compétence spécifique est supposée influencer la réponse à l'item au-delà de l'influence de cette dimension générale (Matteucci & Mignani, 2015).

Ainsi, le modèle *between-item* est particulièrement adéquat pour décrire la performance des étudiants dans chaque domaine évalué ; le modèle *within-item*, quant à lui, permet d'analyser l'influence propre de chacune des dimensions requises à la réussite de l'item (Hartig & Höhler, 2009). Ces deux modèles proposent également différentes définitions et interprétations des échelles de compétences.

L'étude menée par Blömeke *et al.* (2014) sur les données du *Teacher Education and Development Study in Mathematics* (TEDS-M) illustre les différences de perspectives (et leurs implications) de ces deux structures théoriques multidimensionnelles. TEDS-M est une enquête internationale de l'*International Association for the Evaluation of Educational Achievement* (IEA) qui évalue auprès des futurs enseignants deux dimensions des compétences en mathématiques: les connaissances sur les contenus mathématiques (*mathematical contents knowledge*, MCK) et les connaissances sur les contenus pédagogiques propres aux mathématiques (*mathematics pedagogical content knowledge*, MPCK). Ainsi, l'évaluation comporte deux sous-tests : un premier dédié aux MCK et un second centré sur les MPCK. Les auteurs ont calibré les compétences en MCK et en MPCK des futurs enseignants selon deux modèles multidimensionnels. Dans le modèle *between-item*, chaque compétence n'influence que les items propres à cette compétence ; dans le modèle *within-item*, la compétence générale en mathématiques MCK influence tous les items, les items ciblés sur les MPCK mesurant également la compétence relative à la pédagogie des mathématiques. Les résultats de leur analyse des estimations de la performance dans ces deux modèles indiquent que, au niveau pays, les opportunités d'apprentissage des MPCK ne corrèlent pas avec la compétence moyenne en MPCK dans le modèle *between-item* ; par contre, dans le modèle *within-item*, cette corrélation s'élève à 0.30, révélant les bénéfices d'offrir de nombreuses occasions d'enseigner la pédagogie sur la compétence en MPCK des futurs enseignants. Pareillement, les auteurs relèvent que le ranking des pays sur leur moyenne en MPCK fluctue selon le modèle envisagé. Certains pays (dont les USA, la Norvège ou l'Espagne) ont un meilleur classement dans le modèle *within-item*: ces pays mettent l'accent sur la pédagogie lors de la formation initiale, parfois au détriment des contenus mathématiques, ce qui pénalise leur classement MPCK dans le modèle *between-item*. À l'inverse, d'autres pays (dont Taiwan et la Russie) sont moins bien classés avec le modèle *within-item*, leur classement avec le modèle *between-item* étant gonflé par leurs excellentes performances en MCK.

Tel qu'illustré, les enjeux d'une étude de la dimensionnalité d'un test dépassent un intérêt purement psychométrique et peuvent aussi impacter les conclusions pédagogiques découlant de l'analyse des données.

1.2. La modélisation des dimensions dans le cadre des modèles multidimensionnels de réponse à l'item

Les modèles de réponse à l'item (*Item Response Theory*, IRT) postulent que la probabilité de réussir un item est fonction de la compétence[1] de l'élève et d'une ou plusieurs caractéristiques de l'item (Hartig & Höhler, 2009). L'estimation de la compétence des individus dépend donc de la réponse qu'ils ont donnée aux items ainsi que des caractéristiques de ces items (Embretson & Reise, 2000). Les modèles IRT multidimensionnels (MIRT) modélisent la relation entre plusieurs compétences des élèves et la probabilité de réussir un item donné du test (Ackerman, Gierl & Walker, 2003), différentes compétences pouvant influencer la réussite de l'item. Ainsi, les modèles MIRT permettent d'évaluer la structure latente des dimensions qui sous-tend le processus de réponse aux items et amènent une meilleure compréhension des contenus et/ou processus cognitifs mesurés (Ackerman *et al.*, 2003).

Plus spécifiquement, cet article se focalise sur le modèle logistique à 2 paramètres (2PL) compensatoire. Les modèles compensatoires sont fortement associés aux analyses factorielles et sont les plus utilisés dans la littérature (Reckase, 2009). Il s'agit de modèles additifs pour lesquels la probabilité de réussir un item est une combinaison linéaire de compétences (Sijtsma & Junker, 2006). Ainsi, la probabilité de réussite à l'item *i* du sujet *j* est égale à (Reckase, 2009) :

$$P(U_{ij} = 1) = \frac{e^{\boldsymbol{a_i}\boldsymbol{\theta_j'}+d_i}}{1 + e^{\boldsymbol{a_i}\boldsymbol{\theta_j'}+d_i}}$$

Où

- $\boldsymbol{\theta_j}$ est le vecteur des paramètres de compétence du sujet *j*,
- $\boldsymbol{a_i}$ est le vecteur des paramètres de discrimination de l'item *i* et
- d_i, l'*intercept* de l'item *i*, est égal à $-\boldsymbol{a_i}\boldsymbol{b_i'}$ (c'est-à-dire moins une fois le produit du vecteur de discrimination et de la transposée du vecteur des paramètres de difficulté).

L'*intercept* général d_i ne distingue pas explicitement les paramètres de difficulté par dimension. Il fournit une indication de la probabilité globale de réussir l'item *i*: un *intercept* d_i plus élevé traduit une plus grande probabilité de réussite globale, sans pour autant que la probabilité de réussite soit plus élevée pour chaque combinaison de compétences possible, d_i étant fonction du vecteur des paramètres de discrimination (Reckase & McKinley, 1983).

Dans le modèle 2PL, le paramètre de discrimination de l'item *i* pour une dimension *l* donnée, a_{il}, indique l'inclinaison de la pente de la surface de réponse à l'item (*Item Response*

[1] Le terme compétence (en anglais, *ability*) renvoie au positionnement de l'individu sur l'échelle d'une dimension donnée, cette dernière pouvant être d'ordre cognitif (par exemple, la performance en mathématiques) ou non-cognitif (par exemple, la motivation intrinsèque).

Surface) sur cette dimension : lorsque la discrimination augmente, la transition entre une faible probabilité de réussite et une forte probabilité de réussite est plus rapide (Reckase & McKinley, 1983). Le paramètre de discrimination d'un item donné varie d'une dimension à l'autre. Ainsi, le vecteur des paramètres de discrimination d'un item indique l'importance de chacune des dimensions dans la détermination de la probabilité de réussite de cet item (Sheng & Wikle, 2007). Si la discrimination est élevée, alors la dimension correspondante conditionne considérablement la réponse fournie. Par contre, si la discrimination est faible, alors la réponse donnée n'est que très faiblement influencée par cette dimension. Dans un modèle à *m* dimensions, le vecteur des paramètres de discrimination, $\boldsymbol{a_i}$ interagit de manière additive avec le vecteur de compétence, $\boldsymbol{\theta_j}$ (Ackerman *et al.*, 2003 ; Reckase, 2009):

$$\boldsymbol{a_i\theta_j'} = a_{i1}\theta_{j1} + a_{i2}\theta_{j2} + \cdots + a_{im}\theta_{jm} = \sum_{l=1}^{m} a_{il}\theta_{il}$$

C'est de cette additivité du modèle que découle sa propriété compensatoire : une compétence élevée dans une dimension peut compenser une compétence faible sur une autre dimension (Ackerman & Henson, 2015 ; Hartig & Höhler, 2009). La compensation est maximale quand la discrimination est la même à travers les dimensions évaluées par l'item (Ackerman *et al.*, 2003).

L'hypothèse sous-jacente à la propriété compensatoire est que, lorsqu'un étudiant répond à un item, il implique toutes ses connaissances et compétences pour fournir la meilleure réponse possible (Reckase, 2009). Ackerman (1994) donne l'exemple d'un test qui mesurerait deux compétences, à savoir la performance en lecture ainsi que la connaissance d'un sujet particulier (le baseball) : un étudiant faible lecteur pourrait utiliser sa connaissance du baseball afin de répondre correctement à l'item. Cette propriété de compensation n'intervient qu'en cas de multidimensionnalité *within-item* puisque, dans l'autre modèle multidimensionnel, la réponse à un item ne dépend que d'une et une seule dimension. Dans le cas du modèle *within-item* proposé pour les données PISA 2003, l'hypothèse d'additivité des dimensions générale et spécifique apparait raisonnable. Pour l'item sur les taux de change, un étudiant avec une faible compétence en mathématiques (par exemple, une piètre compréhension des contenus mathématiques évalués par l'item) pourrait compenser cette faiblesse par une compétence générale élevée, résultant en une plus grande probabilité de succès.

Les paramètres de discrimination sont donc conditionnés à une dimension particulière ; le paramètre de discrimination multivarié, MDISC, est un indice reflétant la discrimination maximale de l'item qui se calcule comme suit (Ackerman, 1994; Ackerman *et al.*, 2003; Reckase & McKinley, 1991) :

$$MDISC_i = \left(\sum_{l=1}^{m} a_{il}^2\right)^{1/2}$$

Le modèle compensatoire qui vient d'être présenté modélise la probabilité de réussite d'un item dichotomique. Or, certains items de PISA sont des items polytomiques : un élève peut échouer à un tel item, le réussir partiellement (crédit partiel) ou complètement. Le modèle à crédit partiel généralisé (*Generalized Partial Credit Model*, GPCM) de Muraki (1992) permet de modéliser la probabilité de répondre une des catégories ordonnées d'un item polytomique. Dans son extension multidimensionnelle, la probabilité qu'un individu *j* obtienne un score *k* à l'item *i* est égale à (Reckase, 2009) :

$$P(U_{ij} = k) = \frac{e^{k\boldsymbol{a_i\theta'_j} - \sum_{u=0}^{k} \beta_{iu}}}{\sum_{v=0}^{K_i} e^{v\boldsymbol{a_i\theta'_j} - \sum_{u=0}^{v} \beta_{iu}}}$$

où β_{iu} est le paramètre de *threshold* pour la catégorie *u*, β_{i0} étant fixé à 0.

Les trois structures envisagées seront comparées au regard de l'ajustement des modèles. La distribution des paramètres d'items, et plus particulièrement des paramètres de discrimination, sera ensuite détaillée. Plus spécifiquement, dans un modèle *within-item*, ces paramètres reflètent le type de structure sous-jacent à l'évaluation PISA. Si les paramètres de discrimination relatifs à la dimension générale et à la dimension spécifique se dispersent dans l'espace à deux-dimensions formé par ces deux compétences, alors les items mesurent un mélange de compétences et le test a une structure complexe (Ackerman *et al.*, 2003). Au contraire, si tous les items présentent une discrimination élevée sur une dimension et faible voire inexistante sur l'autre, alors le test a une structure simple (proche de celle du modèle unidimensionnel ou *between-item* selon l'axe autour duquel elles se concentrent). Enfin, une structure approximativement simple est une situation intermédiaire où les items sont localisés dans un secteur proche d'une des deux dimensions (Ackerman *et al.*, 2003).

2. Méthodologie

2.1. Données

Les données PISA 2003 sont utilisées dans le cadre de la présente étude. Le choix de ce cycle PISA se justifie par le nombre important d'items du test cognitif rendus publics. Les données de 17 pays (ou Communautés dans le cas de la Belgique: Communauté Flamande et Communauté Française) sont analysées. Ces pays ont été sélectionnés afin de préserver une certaine diversité quant aux contextes éducatifs analysés. Le tableau 1 présente la taille des échantillons pour les pays retenus dans le cadre de cette recherche.

Tableau 1. Taille d'échantillon des pays analysés

Pays	Taille d'échantillon
Australie	6294
Communauté Flamande	5059
Communauté Française	2958
Allemagne	4666
Danemark	4243
Finlande	5796
France	4306
Grande-Bretagne	7072
Hongrie	4802
Japon	4707
Corée	5454
Luxembourg	3927
Pays-Bas	3992
Norvège	4064
Nouvelle-Zélande	4511
Portugal	4608
Suède	4624

2.2. Analyses

Les analyses ne portent que sur les items retenus par l'OCDE lors de la mise à l'échelle internationale des données de 2003 : les items ayant de piètres propriétés psychométriques ont été supprimés soit au niveau national, soit au niveau international (OCDE, 2005). Au final, 165 items sont considérés comme non-problématiques au niveau international. Le domaine principal en 2003 étant les mathématiques, ce domaine comporte plus d'items et le nombre final d'items se répartit comme suit: 84 items de mathématiques, 28 items de lecture, 34 items de sciences et 19 items de résolution de problèmes. Les items supprimés au niveau national ou infranational par l'OCDE (2005) ont également été exclus des analyses. Dans tous les pays, les items non-atteints (les items situés à la fin des livrets qu'un étudiant n'a pas eu le temps d'atteindre) sont considérés comme « non-administrés » (plutôt qu'incorrects) afin de ne pas surestimer la difficulté et la discrimination de ces items dans les présentes analyses, dans la continuité de la méthodologie employée pour PISA 2003.

Les modèles (M)IRT ont été estimés par le logiciel Conquest 3.0 (Adams, Wu & Wilson, 2012). Les analyses ont été réalisées pays par pays, sous l'hypothèse que la distribution de la compétence des individus suit une loi normale de moyenne 0 et d'écart-type 1, cette contrainte étant requise à l'identification des modèles 2PL par Conquest (Adams & Macaskill, 2012). Le modèle 2PL estimé par Conquest 3.0 est une variante du GPCM de Muraki au sein duquel un paramètre de pente est estimé pour chaque catégorie de réponse (Adams & Macaskill, 2012) :

$$P(U_{ij} = k) = \frac{e^{\boldsymbol{a_{ik}\theta'_j} - \sum_{u=0}^{k} \beta_{iu}}}{\sum_{v=0}^{K_i} e^{\boldsymbol{a_{iv}\theta'_j} - \sum_{u=0}^{v} \beta_{iu}}}$$

Étant donné que quatre dimensions sont estimées dans deux des trois modèles investigués, l'estimation du maximum de vraisemblance se base sur une approche de type Monte-Carlo (Reckase, 2009). L'ajustement des différents modèles est comparé sur base du Bayesian Information Criterion (BIC) :

$$BIC = deviance + \kappa * \log(n)$$

Où κ est le nombre de paramètres estimés dans le modèle, *n* est le nombre de sujets et la déviance est une mesure de la qualité de l'ajustement du modèle aux données (liée à la vraisemblance du modèle).

Enfin, la fidélité de l'estimation de la compétence des élèves (EAP/PV *reliability* [2]) indique dans quelle mesure les différents modèles permettent une mesure précise de chaque dimension. Il s'agit d'un indice de fidélité à l'échelle de la population et non de l'individu : le design incomplet de PISA permet des estimations de population précises, ces dernières étant l'objet même de l'enquête, alors que les estimations au niveau individu (c'est-à-dire, l'estimation de la compétence d'un élève donné) sont de piètre précision (Adams, 2005). La fidélité se calcule comme le rapport entre la variance des EAP $var(EAP)$, les EAP étant les estimations de la compétence des élèves, et la variance de la population σ^2 (Adams, 2005) :

$$R = 1 - \frac{\overline{\sigma_P^2}}{\sigma^2} = \frac{\sigma^2 - \overline{\sigma_P^2}}{\sigma^2} = \frac{var(EAP)}{\sigma^2}$$

Cet indice reflète la diminution de l'incertitude autour de l'estimation de la compétence (c'est-à-dire la variance moyenne des distributions postérieures, $\overline{\sigma_P^2}$) permise par la mesure (Adams, 2005).

3. Résultats

3.1. Ajustement des modèles

Le tableau 2 présente l'ajustement des trois modèles dans les différents pays. Plus l'indicateur est petit, meilleur est l'ajustement du modèle aux données. Le modèle unidimensionnel a le moins bon ajustement dans tous les pays. Parmi les modèles multidimensionnels, le modèle *between-item* présente le meilleur ajustement en France et au Portugal tandis que le modèle *within-item* est le mieux ajusté dans les autres pays. Si le gain en matière d'ajustement est élevé lors du passage d'un modèle unidimensionnel vers le modèle multidimensionnel *between-item*, la différence d'ajustement entre le modèle *between-item* et *within-item* est parfois peu substantielle. Ainsi, si le modèle unidimensionnel est clairement le moins bien ajusté aux données, les deux modèles multidimensionnels peuvent être employés afin de fournir une description de l'interaction entre les compétences des élèves et les items de PISA 2003.

Les modèles MIRT *within-item* et *between-item* proposant un meilleur ajustement aux données que le modèle unidimensionnel, la section suivante propose une analyse des paramètres de discrimination de ces modèles, ces paramètres permettant de refléter à quel point les items mesurent chaque dimension et ainsi préciser la structure de l'évaluation PISA 2003.

[2] EAP=*Expected a-posteriori* et PV= *Plausible Value*

Tableau 2. Ajustement (déviance et Bayesian Information Criterion (BIC)) des modèles IRT unidimensionnel (UIRT), MIRT between-item et MIRT within-item. Le modèle ayant le plus petit BIC (et donc s'ajustant le mieux) est, pour chaque pays, indiqué en gras

Pays	Indice d'ajustement	UIRT	*Between-item* MIRT	*Within-item* MIRT	UIRT-*between-item*	*Between-item* - *within-item*
Australie	Déviance	348314	346532	345791		
	BIC	349769	348021	**347812**	1748	209
Com. Flamande	Déviance	275613	274694	273965		
	BIC	277031	276146	**275935**	885	211
Com. Française	Déviance	158073	157143	156621		
	BIC	159403	158503	**158468**	900	35
Allemagne	Déviance	252392	251276	250576		
	BIC	253790	252707	**252516**	1083	191
Danemark	Déviance	233822	232629	231958		
	BIC	235212	234051	**233888**	1161	163
Finlande	Déviance	322660	321492	320627		
	BIC	324101	322967	**322629**	1134	338
France	Déviance	239856	238475	237972		
	BIC	241247	**239899**	239906	1348	-7
Grande-Bretagne	Déviance	391141	389816	388947		
	BIC	392616	391325	**390989**	1291	336
Hongrie	Déviance	252392	251276	250576		
	BIC	253790	252707	**252516**	1083	191
Japon	Déviance	260265	258607	257831		
	BIC	261672	260047	**259785**	1625	262
Corée	Déviance	299686	298520	297677		
	BIC	301103	299970	**299643**	1133	327
Luxembourg	Déviance	218415	217269	216586		
	BIC	219791	218678	**218498**	1113	180
Pays-Bas	Déviance	222530	221894	221344		
	BIC	223909	223306	**223260**	603	46
Norvège	Déviance	222649	221118	220448		
	BIC	224024	222533	**222357**	1491	176
Nouvelle-Zélande	Déviance	248250	247055	246412		
	BIC	249650	248488	**248356**	1162	132
Portugal	Déviance	247790	246802	246381		
	BIC	249186	**248238**	248322	948	-84
Suède	Déviance	254745	253426	252741		
	BIC	256148	254862	**254031**	1286	831

3.2. Paramètres de discrimination des items dans les modèles multidimensionnels

Le tableau 3 présente la moyenne, l'écart-type et le coefficient de variation des paramètres de discrimination des items, par sous-test, dans les deux modèles multidimensionnels.

En moyenne, la discrimination des items sur la dimension générale dans le modèle *within-item* est proche de celle estimée en *between-item*. Les discriminations relatives aux compétences spécifiques en mathématiques, lecture et sciences (modèle *within-item*) sont,

par contre, nettement plus faibles. Le modèle *within-item* indique donc que les items fournissent une meilleure mesure de la compétence générale que des compétences spécifiques : en moyenne, les items permettent de mieux distinguer deux élèves de compétence générale différente que deux élèves de compétence spécifique différente. En d'autres termes, dans le modèle *within-item*, la compétence générale des élèves est, en moyenne, plus déterminante de leur probabilité de réussite que leurs compétences spécifiques en mathématiques, lecture ou sciences.

Tableau 3. Moyenne μ, écart-type σ et coefficient de variation cv de la distribution des paramètres de discrimination des items, par sous-test, dans les modèles between-item et within-item

		Discrimination dans les modèles MIRT :								
					Within-item					
		Between-item			**Gen.**			**Spe.**		
Pays	**Sous-test**	μ	σ	cv	μ	σ	cv	μ	σ	cv
Australie	Mathématiques	1.523	0.662	43.430	1.437	0.635	44.215	0.446	0.488	109.339
	Lecture	1.498	0.364	24.290	1.340	0.345	25.761	0.793	0.268	33.841
	Sciences	1.336	0.546	40.885	1.182	0.502	42.471	0.533	0.383	71.808
	Résolution de problèmes	1.620	0.613	37.820	1.485	0.587	39.534			
Communauté Flamande	Mathématiques	1.499	0.604	40.334	1.488	0.610	40.957	0.270	0.499	184.517
	Lecture	1.446	0.472	32.648	1.313	0.456	34.753	0.699	0.462	66.050
	Sciences	1.272	0.602	47.356	1.164	0.551	47.304	0.549	0.371	67.669
	Résolution de problèmes	1.579	0.616	39.018	1.508	0.584	38.686			
Communauté Française	Mathématiques	1.518	0.665	43.812	1.419	0.649	45.715	0.659	0.601	91.314
	Lecture	1.742	0.501	28.746	1.747	0.650	37.204	0.497	0.783	157.622
	Sciences	1.384	0.688	49.740	1.312	0.651	49.649	0.540	0.607	112.431
	Résolution de problèmes	1.691	0.642	37.990	1.633	0.637	39.041			
Allemagne	Mathématiques	1.513	0.539	35.651	1.525	0.574	37.596	0.240	0.446	186.105
	Lecture	1.554	0.469	30.198	1.404	0.460	32.762	0.755	0.443	58.607
	Sciences	1.450	0.659	45.481	1.351	0.638	47.210	0.610	0.436	71.440
	Résolution de problèmes	1.620	0.645	39.811	1.524	0.587	38.550			
Danemark	Mathématiques	1.371	0.615	44.841	1.362	0.711	52.200	0.466	0.681	146.283
	Lecture	1.272	0.391	30.707	1.050	0.337	32.128	0.763	0.286	37.519
	Sciences	1.259	0.610	48.468	1.166	0.539	46.190	0.525	0.482	91.883
	Résolution de problèmes	1.479	0.603	40.791	1.410	0.572	40.534			
Finlande	Mathématiques	1.226	0.537	43.794	1.260	0.666	52.856	0.143	0.596	418.179
	Lecture	1.268	0.378	29.808	1.034	0.345	33.311	0.755	0.342	45.368
	Sciences	1.114	0.461	41.437	1.012	0.416	41.046	0.514	0.418	81.464
	Résolution de problèmes	1.392	0.534	38.364	1.319	0.504	38.209			
France	Mathématiques	1.322	0.546	41.282	1.300	0.554	42.606	0.311	0.472	151.955
	Lecture	1.471	0.379	25.753	1.234	0.344	27.866	0.817	0.366	44.760
	Sciences	1.334	0.585	43.867	1.194	0.512	42.846	0.633	0.430	67.945
	Résolution de problèmes	1.460	0.553	37.885	1.361	0.513	37.706			
Grande-Bretagne	Mathématiques	1.517	0.785	51.738	1.524	0.786	51.572	0.001	0.424	31875.119
	Lecture	1.535	0.400	26.038	1.327	0.382	28.761	0.697	0.334	47.906
	Sciences	1.374	0.606	44.065	1.247	0.550	44.108	0.563	0.313	55.606
	Résolution de problèmes	1.600	0.630	39.392	1.473	0.580	39.384			
Hongrie	Mathématiques	1.432	0.757	52.823	1.428	0.840	58.830	0.496	0.718	144.672
	Lecture	1.302	0.519	39.810	1.164	0.539	46.280	0.755	0.804	106.489
	Sciences	1.134	0.498	43.928	1.007	0.440	43.711	0.550	0.372	67.520
	Résolution de problèmes	1.584	0.718	45.361	1.517	0.680	44.798			
Japon	Mathématiques	1.480	0.610	41.203	1.420	0.695	48.937	0.575	0.685	119.238
	Lecture	1.444	0.483	33.462	1.265	0.437	34.561	0.793	0.482	60.741
	Sciences	1.325	0.676	51.035	1.198	0.591	49.365	0.604	0.489	80.964
	Résolution de problèmes	1.639	0.661	40.304	1.564	0.628	40.134			

		Discrimination dans les modèles MIRT :								
					Within-item					
		Between-item			Gen.			Spe.		
Pays	**Sous-test**	μ	σ	cv	μ	σ	cv	μ	σ	cv
Corée	Mathématiques	1.428	0.738	51.630	1.484	0.857	57.746	0.052	0.566	1086.807
	Lecture	1.235	0.382	30.888	1.047	0.331	31.626	0.689	0.291	42.277
	Sciences	1.261	0.509	40.319	1.149	0.456	39.663	0.572	0.428	74.966
	Résolution de problèmes	1.452	0.681	46.910	1.332	0.580	43.556			
Luxembourg	Mathématiques	1.415	0.579	40.892	1.377	0.595	43.177	0.471	0.572	121.534
	Lecture	1.474	0.430	29.197	1.314	0.397	30.200	0.706	0.502	71.031
	Sciences	1.362	0.731	53.617	1.292	0.749	57.963	0.601	0.583	96.868
	Résolution de problèmes	1.611	0.663	41.163	1.517	0.605	39.893			
Pays-Bas	Mathématiques	1.394	0.656	47.077	1.447	0.781	53.991	0.080	0.571	712.181
	Lecture	1.415	0.424	29.938	1.272	0.408	32.069	0.643	0.447	69.593
	Sciences	1.256	0.538	42.844	1.155	0.501	43.347	0.496	0.250	50.297
	Résolution de problèmes	1.538	0.659	42.820	1.446	0.619	42.803			
Norvège	Mathématiques	1.419	0.573	40.373	1.453	0.717	49.343	0.261	0.726	278.710
	Lecture	1.519	0.420	27.682	1.269	0.376	29.642	0.900	0.390	43.350
	Sciences	1.325	0.680	51.321	1.167	0.580	49.734	0.665	0.455	68.430
	Résolution de problèmes	1.600	0.585	36.576	1.504	0.562	37.360			
Nouvelle-Zélande	Mathématiques	1.494	0.624	41.791	1.463	0.655	44.758	0.452	0.586	129.789
	Lecture	1.624	0.413	25.425	1.467	0.390	26.585	0.787	0.420	53.433
	Sciences	1.302	0.539	41.437	1.194	0.492	41.169	0.535	0.401	74.868
	Résolution de problèmes	1.547	0.612	39.563	1.480	0.586	39.625			
Portugal	Mathématiques	1.430	0.759	53.113	1.362	0.729	53.492	0.499	0.469	93.947
	Lecture	1.399	0.486	34.732	1.248	0.475	38.035	0.679	0.375	55.170
	Sciences	1.147	0.542	47.265	1.064	0.500	47.027	0.436	0.309	70.827
	Résolution de problèmes	1.471	0.618	42.024	1.405	0.588	41.828			
Suède	Mathématiques	1.462	0.706	48.241	1.406	0.763	54.272	0.578	0.697	120.660
	Lecture	1.417	0.429	30.251	1.261	0.430	34.135	0.756	0.634	83.798
	Sciences	1.300	0.624	47.963	1.228	0.574	46.717	0.509	0.514	100.951
	Résolution de problèmes	1.479	0.610	41.256	1.420	0.599	42.184			

Les paramètres de discrimination de cette dimension générale corrèlent fortement avec les discriminations des compétences estimées dans le modèle *between-item* (tableau A1, en annexe). Ainsi, les paramètres de discrimination relatifs à la dimension générale présentent une corrélation supérieure à 0.91 avec les paramètres de discrimination estimés dans le modèle *between-item* dans tous les pays et pour tous les sous-tests, à l'exception du sous-test en compréhension de l'écrit en Communauté Française de Belgique (corrélation égale à 0.772). Ces résultats indiquent que les dimensions estimées dans une modélisation *between-item* correspondent principalement à la dimension générale estimée dans un modèle *within-item*.

Les corrélations entre les discriminations des dimensions spécifiques et celles de la dimension générale (tableau A1) varient d'un pays à l'autre et d'un sous-test à l'autre. Si elles tendent à être positives, surtout pour le sous-test de sciences, ces corrélations sont instables et peuvent être fortement influencées par des valeurs extrêmes. À titre d'exemple, la figure 3 présente les diagrammes de dispersion des paramètres de discrimination du modèle *within-item*, pour la Communauté Française de Belgique, la discrimination de la compétence spécifique étant en abscisse et celle de la dimension générale en ordonnée. Les trois premiers graphiques concernent chacun un sous-test (figure 3a : mathématiques, 3b :

lecture et 3c : sciences) ; ils sont accompagnés d'une boite à moustache[3] présentant la dispersion des paramètres de discrimination de la compétence spécifique associée. Le quatrième diagramme de dispersion (figure 3d) regroupe l'ensemble des items des trois sous-tests. Il s'accompagne d'un gradient de couleur correspondant à la discrimination de ces items dans le modèle *between-item*. Pour le sous-test en lecture (figure 3b), il y a deux valeurs extrêmes pour le paramètre de discrimination de la compétence spécifique (entourées) : ces deux items présentent une discrimination extrêmement faible de la compétence spécifique en lecture et très élevée pour la dimension générale. Sans ces deux items, la corrélation entre la discrimination de la dimension générale et la discrimination de la compétence spécifique en lecture passe de -0.222 à 0.467. La relativement faible corrélation entre la discrimination de la dimension générale et la discrimination estimée dans le modèle *between-item* en Communauté Française pour la lecture s'explique également par ces deux items : après suppression de ces items, la corrélation passe de 0.772 à 0.952.

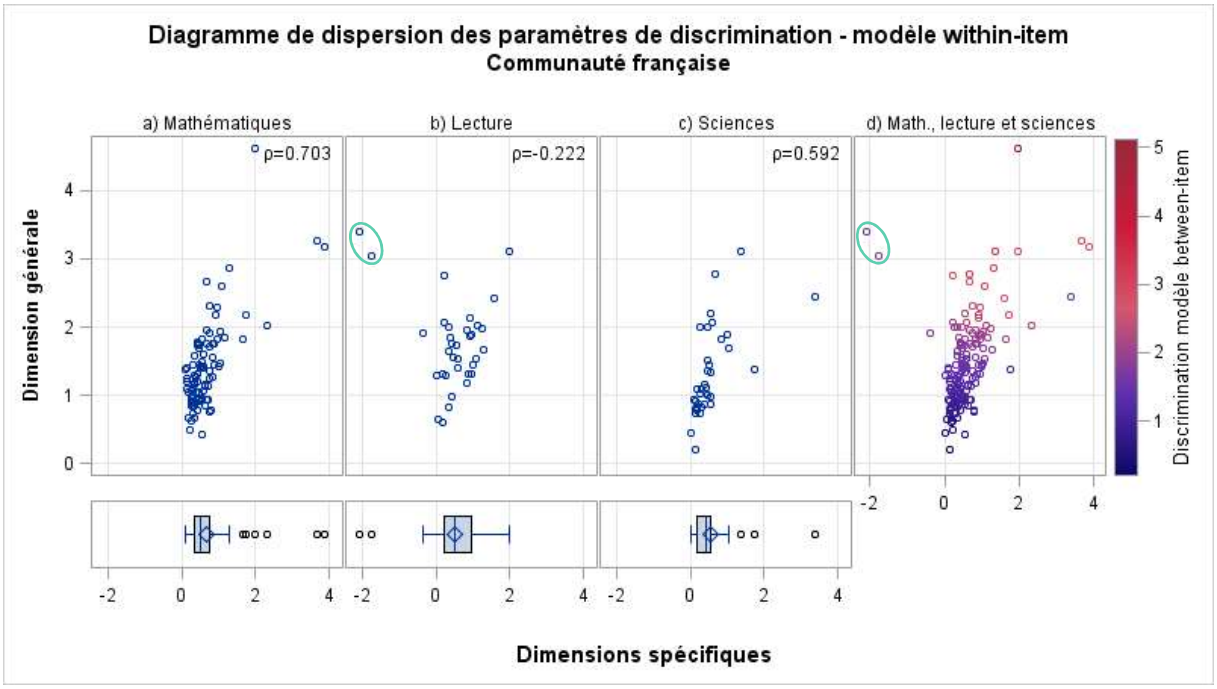

Figure 3. Pour le modèle *within-item* de la Communauté Française de Belgique, diagramme de dispersion des paramètres de discrimination des items a) de mathématiques, b) de lecture, c) de sciences et d) de ces 3 domaines envisagés simultanément. Les boites à moustache présentent la distribution de la compétence spécifique en mathématiques, lecture et sciences. Les deux items de lecture entourés présentent des valeurs extrêmes. Les discriminations des crédits partiels sont inclues.

Ainsi, ce n'est pas parce que la discrimination d'un item est élevée sur, par exemple, la dimension générale qu'elle le sera aussi nécessairement sur la dimension spécifique : la dimension spécifique peut donc affecter différemment les items. De plus, la discrimination des dimensions spécifiques est, en moyenne, plus faible que celle de la dimension générale. Ces résultats traduisent une structure approximativement simple dans la majorité des sous-

[3] La boite (le rectangle) s'étend du premier quartile (extrémité gauche de la boite) au troisième quartile (extrémité droite) ; le médian est représenté par la droite centrale et la moyenne par un losange. Les moustaches s'étendent jusqu'à la plus petite valeur (à gauche) ou plus grande valeur (à droite) observée dans une limite de 1.5 fois l'écart interquartile au-delà de la « boite ».

tests : tel qu'illustré pour la Communauté Française dans la figure 3, les paramètres de discrimination des items se dispersent principalement autour de l'axe des ordonnées, c'est-à-dire de la dimension générale. En d'autres termes, le modèle *within-item* indique que les items mesurent un mélange d'une dimension générale et d'une dimension spécifique au domaine envisagé, la première étant nettement plus déterminante de la réussite que la seconde. Les paramètres estimés dans le modèle *within-item* pour l'item 11 de l'unité « taux de change » permettent d'illustrer l'interaction des dimensions générale et spécifique et, plus spécifiquement, les implications de la meilleure discrimination de la dimension générale.

3.3. Modélisation de la probabilité de réussir l'item « Taux de change » (modélisation MIRT within-item)

La figure 4a présente l'*Item Response Surface* de l'item « taux de change », toujours en Communauté Française de Belgique. Les paramètres estimés pour cet item sont l'*intercept* et la discrimination sur les dimensions spécifique en mathématiques et générale, les discriminations sur les dimensions spécifiques en lecture et en sciences étant, de fait, fixées à 0. Cette surface représente la probabilité de répondre correctement à l'item pour chaque combinaison de compétences générale et spécifique. Pour un élève *j* de compétence sur la dimension spécifique en mathématiques θ_{1j}, en lecture θ_{2j} et en sciences θ_{3j} et sur la dimension générale θ_{4j}, elle se calcule par le biais de la formule générale du modèle 2PL pour les paramètres d'item estimés :

$$P(U_{ij}=1)=\frac{e^{\boldsymbol{a_i\theta'_j}+d_i}}{1+e^{\boldsymbol{a_i\theta'_j}+d_i}}=\frac{e^{(0.891*\theta_{1j}+0*\theta_{2j}+0*\theta_{3j}+1.436*\theta_{4j}-0.039)}}{1+e^{(0.891*\theta_{1j}+0*\theta_{2j}+0*\theta_{3j}+1.436*\theta_{4j}-0.039)}}$$

La pente de cette surface est plus forte pour la dimension générale que pour la dimension spécifique. Prenons deux élèves peu performants en mathématiques (par exemple, avec une compétence de -3) mais avec un niveau différent de compétence sur la dimension générale ; le premier, faible également sur la compétence générale (ex : -3) aura une probabilité quasiment nulle de réussir l'item tandis que le second, de compétence générale égale à +3, aura une probabilité de réussir l'item égale à 0.831. Inversement, pour deux élèves de faible compétence générale (ex : -3), un élève avec une compétence spécifique en mathématiques de +3 aura une probabilité de réussite égale à 0.158 alors qu'un élève faible également sur cette dimension n'aura presqu'aucune chance de réussir l'item. Ainsi, si une compétence élevée dans une dimension peut effectivement compenser une faible compétence sur l'autre dimension, cette compensation est relativement limitée et l'influence de la dimension générale sur la probabilité de réussir l'item est nettement plus forte que celle de la dimension spécifique en mathématiques. À l'inverse, si cet item avait la même discrimination sur les deux dimensions, la compensation serait maximale. Cette situation fictive est illustrée dans la figure 4b. La pente de la surface sur les deux dimensions est identique. Dans cette situation, une augmentation de la dimension générale donne lieu à un même accroissement de la probabilité de réussite de l'item qu'une augmentation de même ampleur de la dimension spécifique.

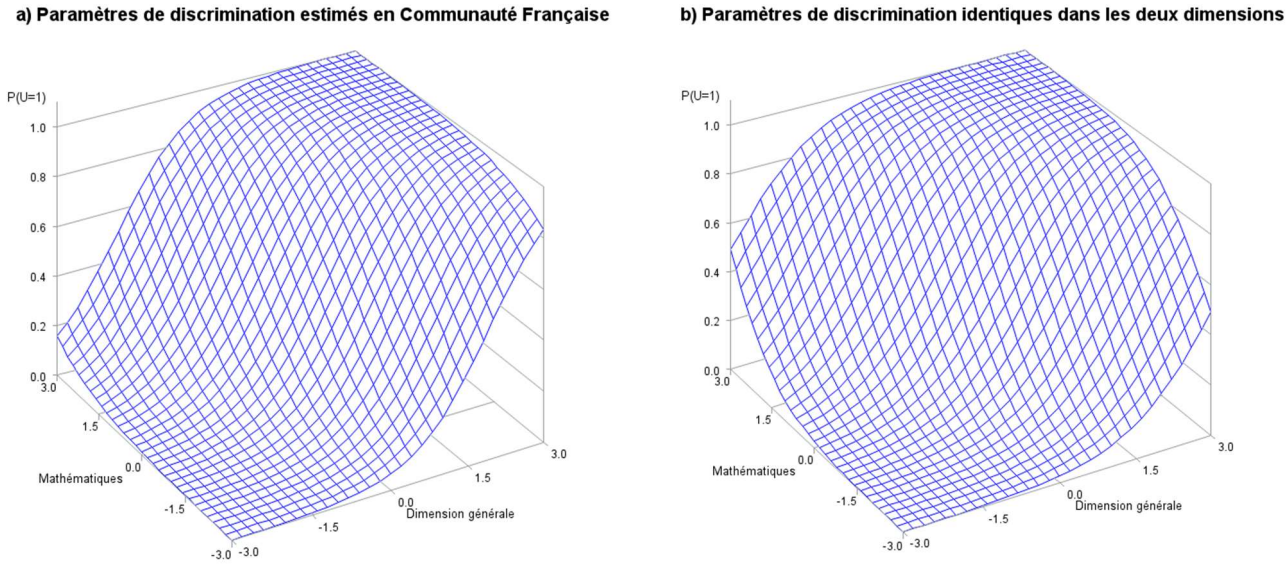

Figure 4. *Item Response Surface* de l'item « taux de change » a) en Communauté Française de Belgique et b) si la discrimination de la compétence générale et de la compétence spécifique étaient identiques. Les paramètres d'item sont (a) $\boldsymbol{d_i}$= -0.039, $\boldsymbol{a_i} = [\mathbf{0.891} \quad \mathbf{0}^* \quad \mathbf{0}^* \quad \mathbf{1,436}]$ et MDISC=1.690 et (b) $\boldsymbol{d_i}$= -0.039, $\boldsymbol{a_i} = [\mathbf{1.195} \quad \mathbf{0}^* \quad \mathbf{0}^* \quad \mathbf{1.195}]$ et MDISC=1.690

Ces différences de discrimination des items impliquent donc qu'ils fournissent une meilleure mesure de la dimension générale que des dimensions spécifiques, ce qui n'est pas sans répercussion sur la précision de l'estimation des compétences dans ce modèle.

3.4. Fidélité des compétences estimées

La fidélité de la compétence estimée est élevée dans les modèles unidimensionnel et multidimensionnel *between-item* (le tableau A2, en annexe, présente l'estimation de la fidélité de la/des compétence(s) estimée(s) dans les trois modèles investigués). En moyenne, la fidélité est égale à 0.926 dans le modèle unidimensionnel et, dans le modèle multidimensionnel *between-item*, à 0.896 pour la compétence en mathématiques, 0.780 en lecture, 0.805 en sciences et 0.832 en résolution de problèmes.

Dans le modèle *within-item*, seule la dimension générale est estimée avec précision. Les fidélités moyennes sont égales à 0.369, 0.252 et 0.230 pour les compétences spécifiques en mathématiques, lecture et sciences et à 0.898 pour la dimension générale. Dans ce modèle, les fidélités estimées pour les trois dimensions spécifiques sont très faibles, dénotant une forte imprécision de leur mesure, en particulier pour les dimensions lecture et sciences qui comportent moins d'items. Cette moindre fidélité des dimensions spécifiques en comparaison de la fidélité de la dimension générale est consistante avec l'analyse des paramètres de discrimination des items réalisée au point 3.2. Étant donné que, dans le modèle *within-item*, les items discriminent, en moyenne, fortement la dimension générale et faiblement la dimension spécifique, cette dernière est estimée avec moins de précision que la première.

4. Perspectives et limitations

Les résultats de cette étude indiquent que les deux modèles multidimensionnels (*between-item* MIRT et *within-item* MIRT) s'ajustent mieux aux données qu'un modèle unidimensionnel. Le modèle *between-item*, qui propose une structure simple, fournit une estimation précise des compétences en mathématiques, lecture, sciences et résolution de problèmes, ce qui se traduit par des paramètres discrimination d'items élevés et une bonne fidélité des estimations des compétences. Le modèle *within-item* postule une structure complexe des dimensions et permet de décrire comment les compétences générale et spécifique influencent la réussite des items. Il apparait que c'est principalement la dimension générale qui est déterminante de la probabilité de réussite d'un item, la dimension spécifique ayant une importance souvent moindre. La fidélité de ces compétences spécifiques est, d'ailleurs, très faible, indiquant un faible degré de précision de ces mesures dans le modèle *within-item*. Ces compétences spécifiques doivent cependant être inclues dans le modèle, le modèle unidimensionnel ayant le moins bon ajustement aux données. À l'inverse, la dimension générale possède une bonne fidélité. La discrimination de cette dimension dans les différents sous-tests est plus élevée et corrèle très fortement avec les discriminations estimées en *between-item*, indiquant que les dimensions du modèle *between-item* sont constituées principalement de cette composante générale ; ce résultat est concordant avec les fortes corrélations latentes observées dans le modèle *between-item*.

L'estimation de ce modèle multidimensionnel *within-item* amène des éléments de compréhension quant au contenu de l'évaluation PISA 2003 dans certaines limites. D'une part, l'interprétation des dimensions du modèle *within-item* n'est pas unique. Comment définir cette dimension générale et ces dimensions spécifiques ? Quatre hypothèses peuvent être proposées :

- La dimension générale est une compétence de résolution de problèmes ou de littéracie tandis que les compétences spécifiques relèvent de compétences et connaissances propres à un domaine ;
- Les dimensions générales et spécifiques correspondent à différentes formes d'intelligence. Par exemple, Brunner (2008) a testé plusieurs modèles structuraux des habiletés cognitives, proposant à la fois des structures simples et complexes, sur les données allemandes de PISA 2000, ces dernières comportant à la fois le test cognitif PISA et un test d'intelligence ;
- La dimension générale n'est pas exclusivement de nature cognitive (contrairement aux deux premières hypothèses) mais renvoie aussi à une dimension motivationnelle : PISA étant une épreuve à faible enjeu pour les élèves, la motivation et la résistance à la fatigue de ces derniers jouent un rôle prépondérant dans le bon déroulement de l'épreuve. La dimension générale serait donc en partie non-cognitive tandis que les dimensions spécifiques renvoient à des dimensions cognitives propres aux domaines ;
- Les dimensions spécifiques sont en partie des artéfacts méthodologiques qui reflètent des différences entre les items de différents domaines. Au-delà de solliciter des connaissances et compétences différentes, les items de différents domaines sont présentés différemment et peuvent solliciter d'autres modalités de réponse ; les compétences spécifiques peuvent donc, en partie, renvoyer à des différences dans les modalités de présentation et les formats de réponse des items.

D'autre part, et indépendamment de leur définition en des termes cognitifs, non cognitifs ou artéfactuels, les dimensions du modèle *between-item* sont plus facilement interprétables. Contrairement au modèle *within-item*, le modèle multidimensionnel *between-item* permet de décrire les différentes échelles de compétence par rapport au contenu des items. Chaque item ne mesurant qu'une et une seule dimension, celle-ci peut donc être décrite en présentant les items qui la mesurent : c'est d'ailleurs ce que l'OCDE fait à travers la publication d'une partie des items. Par ailleurs, le modèle *between-item* est adéquat pour présenter les résultats généraux de l'étude PISA ; par contre, le modèle *within-item* peut être employé pour des analyses approfondies de l'équité et de l'efficacité des systèmes éducatifs. Par exemple, concernant les différences de genre, les écarts en mathématiques dans les études PISA sont minimes. Avec un modèle *within-item*, il serait possible d'analyser cette différence tant sur la dimension générale que sur la compétence propre aux mathématiques.

Une autre application de ce modèle pourrait porter sur les *dodgy items*, ces items supprimés dans un ou plusieurs pays en raison de leurs piètres propriétés psychométriques. Si ces mauvaises propriétés psychométriques ne résultent pas d'une erreur de traduction ou d'impression, un modèle *within-item* pourrait permettre de mieux comprendre ce que mesurent réellement ces items.

Enfin, si ce modèle *within-item* offre des perspectives en termes d'approfondissement des résultats PISA, une de ses limites tient à l'impossibilité de dégager des indicateurs de tendance. En effet, le contenu de la dimension générale varierait au fil des cycles de par la composition des épreuves PISA : si, en 2003, l'évaluation cognitive comprenait des items de résolution de problèmes, d'autres cycles ne comprennent plus ces items mais abordent d'autres domaines, ce qui peut influencer la signification de la dimension générale. Pareillement, les proportions d'items en mathématiques, lecture et sciences varient en fonction des cycles, le domaine majeur contenant le plus d'items. L'informatisation des tests devrait cependant permettre de multiplier les livrets et ainsi de réduire les différences d'importance entre les domaines majeurs et mineurs, résolvant partiellement ce problème.

Comme on peut le constater, cette recherche n'a pas pour ambition de suggérer aux responsables de l'enquête PISA un modèle alternatif pour la mise à l'échelle des données cognitives. En effet, les indicateurs de tendance sont politiquement très importants et il importe de limiter les risques qui pourraient les invalider. Toutefois, ce modèle pourrait nous aider à mieux décrire les compétences sollicitées par chacun des items et éventuellement, à l'instar de l'étude menée par Blömeke *et al.* (2014) sur les données TEDS-M, d'identifier des liens entre certaines compétences spécifiques et des variables contextuelles.

5. Bibliographie

Ackerman, T. A. (1994). Using multidimensional Item Response Theory to understand what items and tests are measuring. *Applied measurement in education, 7*, 225-278.

Ackerman, T. A., Gierl, M. J., & Walker, C. M. (2003). Using multidimensional item response theory to evaluate educational and psychological tests. *Educational measurement: issues and practice, 22*, 37–51.

Ackerman, T. A., & Henson, R. A. (2015). Graphical representations of items and tests that are measuring multiple abilities. In R.E. Millsap, D. M. Bolt, L. A. van der Ark & W.-C. Wang, *Quantitative psychology research* (pp. 113-132). New-York, NY: Springer.

Adams, R. J. (2005). Reliability as a measurement design effect. *Studies in educational evaluation, 31*, 162-172.

Adams, R., & Macaskill, G. (2012). *Score estimation and generalised partial credit models*. Retrieved from https://www.acer.org/conquest/notes-tutorials

Adams, R., Wu, M., & Wilson, M. (2012). *ACER ConQuest 3.0. (computer program)*. Melbourne, Australie: ACER.

Blömeke, S., Houang, R. T., & Suhl, U. (2014) Diagnosing teacher knowledge by applying multidimensional Item Response theory and Multiple-group models. In S. Blömeke, F.-J. Hsieh, G. Kaiser, & W. Schmidt, *International perspectives on teacher knowledge, beliefs and opportunities to learn: TEDS-M results* (pp. 483-501). Dordrecht, The Netherlands: Springer.

Brunner, M. (2008). No *g* in education? *Learning and individual differences, 18,* 152-165.

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

Hartig, J., & Höhler, J. (2009). Multidimensional IRT models for the assessment of competencies. *Studies in Educational Evaluation, 35*, 57-63.

Matteucci, M., & Mignani, S. (2015). Multidimensional IRT models to analyze learning outcomes of Italian students at the end of lower secondary school. In R.E. Millsap, D. M. Bolt, L. A. van der Ark, & W.-C. Wang (Eds.), *Quantitative psychology research* (pp. 91-111). New York, NY: Springer.

Muraki, E. (1992). *A Generalized Partial Credit Model: Application of an EM algorithm*. Princeton, New Jersey: Educational Testing Service.

OCDE. (2004). *Learning for tomorrow's world: first results from PISA 2003*. Paris, France: OECD publishing.

OCDE. (2005). *PISA 2003 : technical report*. Paris, France: OECD Publishing.

OCDE. (2016). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic and Financial Literacy*. Paris, France: OECD Publishing, Paris.

Reckase, M. D. (2009). *Multidimensional Item Response Theory*. New-York, NY: Springer

Reckase, M. D., & McKinley, R. L. (1983, April). *The definition of difficulty and discrimination for multidimensional item response theory models*. Paper presented at the annual meeting of the American Educational Research Association, Quebec, Canada.

Reckase, M. D., & McKinley, R. L. (1991). The discrimination power of items that measure more than one dimension. *Applied psychological measurement, 15*, 361-373.

Sheng, Y., & Wikle, C. K. (2007). Comparing multiunidimensional and unidimensional item response theory models. *Educational and psychological measurement, 67*, 899-919.

Sijtsma, K., & Junker, B. W. (2006). Item Response Theory: past performance, present developments, and future expectations. *Behaviormetrika, 33*, 75-102.

Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological methods, 9*, 116-136.

Wang, W.-C., Wilson, M., & Adams, R. J. (1997). Rasch models for multidimensionality between-items and within-items. In M. Wilson, G. Engelhard, & K. Draney, *Objective measurement: theory into practice, volume 4* (pp. 139-155). Greenwich, CT: Ablex.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis. Rasch measurement.* Chicago, IL: Mesa.

Wu, M., & Adams, R. (2006). Modelling mathematics problem solving item responses using a multidimensional IRT model. *Mathematics education research journal, 18,* 93-113.

6. Annexes

Tableau A1. Corrélation entre les estimations des paramètres de discrimination dans les modèles multidimensionnels between-item et within-item par sous-test

		Corrélations entre la discrimination des items dans les modèles MIRT :		
Pays	**Sous-test**	***Between-item* et comp. générale (*within-item*)**	***Between-item* et comp. spécifique (*within-item*)**	**Comp. générale et spécifique (*within-item*)**
Australie	Mathématiques	0.986	0.638	0.594
	Lecture	0.986	0.495	0.386
	Sciences	0.990	0.810	0.807
	Résolution de problèmes	0.993		
Communauté Flamande	Mathématiques	0.982	0.358	0.302
	Lecture	0.980	0.419	0.518
	Sciences	0.998	0.747	0.727
	Résolution de problèmes	0.998		
Communauté Française	Mathématiques	0.984	0.658	0.703
	Lecture	0.772	0.302	-0.222
	Sciences	0.964	0.399	0.592
	Résolution de problèmes	0.998		
Allemagne	Mathématiques	0.990	0.012	-0.039
	Lecture	0.985	0.240	0.239
	Sciences	0.996	0.911	0.918
	Résolution de problèmes	0.998		
Danemark	Mathématiques	0.950	0.560	0.691
	Lecture	0.982	0.510	0.367
	Sciences	0.992	0.610	0.626
	Résolution de problèmes	0.998		
Finlande	Mathématiques	0.937	0.339	0.446
	Lecture	0.981	0.417	0.258
	Sciences	0.991	0.492	0.502
	Résolution de problèmes	0.997		
France	Mathématiques	0.993	0.022	0.038
	Lecture	0.974	0.588	0.424
	Sciences	0.994	0.740	0.732
	Résolution de problèmes	0.997		
Grande-Bretagne	Mathématiques	0.994	0.090	0.100
	Lecture	0.981	0.310	0.197
	Sciences	0.998	0.844	0.825
	Résolution de problèmes	0.998		
Hongrie	Mathématiques	0.985	0.794	0.827
	Lecture	0.913	0.236	0.529
	Sciences	0.990	0.820	0.753
	Résolution de problèmes	0.999		
Japon	Mathématiques	0.915	0.550	0.756
	Lecture	0.981	0.510	0.425
	Sciences	0.995	0.848	0.829
	Résolution de problèmes	0.996		
Corée	Mathématiques	0.977	-0.200	-0.140
	Lecture	0.986	0.333	0.218
	Sciences	0.990	0.638	0.609
	Résolution de problèmes	0.997		
Luxembourg	Mathématiques	0.975	0.449	0.529
	Lecture	0.977	0.202	0.257
	Sciences	0.993	0.920	0.923
	Résolution de problèmes	0.994		

		Corrélations entre la discrimination des items dans les modèles MIRT :		
Pays	**Sous-test**	***Between-item* et comp. générale (*within-item*)**	***Between-item* et comp. spécifique (*within-item*)**	**Comp. générale et spécifique (*within-item*)**
Pays-Bas	Mathématiques	0.969	-0.325	-0.406
	Lecture	0.982	0.337	0.389
	Sciences	0.998	0.710	0.691
	Résolution de problèmes	0.999		
Norvège	Mathématiques	0.918	-0.081	0.080
	Lecture	0.979	0.516	0.361
	Sciences	0.997	0.890	0.868
	Résolution de problèmes	0.995		
Nouvelle-Zélande	Mathématiques	0.990	0.642	0.629
	Lecture	0.986	0.490	0.438
	Sciences	0.993	0.583	0.610
	Résolution de problèmes	0.999		
Portugal	Mathématiques	0.996	0.693	0.652
	Lecture	0.987	0.130	0.045
	Sciences	0.997	0.754	0.714
	Résolution de problèmes	0.999		
Suède	Mathématiques	0.988	0.786	0.797
	Lecture	0.953	0.492	0.589
	Sciences	0.983	0.483	0.581
	Résolution de problèmes	0.993		

Tableau A2. Fidélité des estimations EAP/PV

		Fidélité de l'estimation de la performance EAP/PV dans les modèles (M)IRT		
Pays	**Dimension**	**Unidimensionnel**	***Between-item***	***Within-item***
Australie	Mathématiques		0.875	0.375
	Lecture		0.774	0.247
	Sciences		0.795	0.211
	Résolution de problèmes		0.804	
	Générale	0.932		0.903
Communauté Flamande	Mathématiques		0.912	0.351
	Lecture		0.802	0.216
	Sciences		0.819	0.214
	Résolution de problèmes		0.848	
	Générale	0.933		0.899
Communauté Française	Mathématiques		0.907	0.435
	Lecture		0.804	0.260
	Sciences		0.808	0.231
	Résolution de problèmes		0.851	
	Générale	0.931		0.892
Allemagne	Mathématiques		0.912	0.362
	Lecture		0.822	0.256
	Sciences		0.837	0.230
	Résolution de problèmes		0.862	
	Générale	0.938		0.918
Danemark	Mathématiques		0.889	0.381
	Lecture		0.738	0.261
	Sciences		0.787	0.246
	Résolution de problèmes		0.831	
	Générale	0.922		0.881
Finlande	Mathématiques		0.880	0.373
	Lecture		0.745	0.251
	Sciences		0.786	0.236
	Résolution de problèmes		0.810	
	Générale	0.911		0.892
France	Mathématiques		0.886	0.369
	Lecture		0.750	0.267
	Sciences		0.783	0.248
	Résolution de problèmes		0.810	
	Générale	0.918		0.891
Grande-Bretagne	Mathématiques		0.904	0.317
	Lecture		0.798	0.242
	Sciences		0.822	0.223
	Résolution de problèmes		0.833	
	Générale	0.931		0.923
Hongrie	Mathématiques		0.888	0.358
	Lecture		0.771	0.265
	Sciences		0.777	0.239
	Résolution de problèmes		0.842	
	Générale	0.917		0.883
Japon	Mathématiques		0.898	0.404
	Lecture		0.769	0.268
	Sciences		0.787	0.243
	Résolution de problèmes		0.809	
	Générale	0.929		0.886

		Fidélité de l'estimation de la performance EAP/PV dans les modèles (M)IRT		
Pays	**Dimension**	**Unidimensionnel**	***Between-item***	***Within-item***
Corée	Mathématiques		0.900	0.326
	Lecture		0.756	0.243
	Sciences		0.802	0.259
	Résolution de problèmes		0.823	
	Générale	0.925		0.902
Luxembourg	Mathématiques		0.897	0.378
	Lecture		0.786	0.256
	Sciences		0.811	0.230
	Résolution de problèmes		0.826	
	Générale	0.928		0.892
Pays-Bas	Mathématiques		0.909	0.349
	Lecture		0.827	0.227
	Sciences		0.847	0.206
	Résolution de problèmes		0.868	
	Générale	0.928		0.923
Norvège	Mathématiques		0.892	0.375
	Lecture		0.759	0.293
	Sciences		0.788	0.262
	Résolution de problèmes		0.824	
	Générale	0.926		0.902
Nouvelle-Zélande	Mathématiques		0.906	0.373
	Lecture		0.806	0.244
	Sciences		0.819	0.232
	Résolution de problèmes		0.853	
	Générale	0.931		0.910
Portugal	Mathématiques		0.891	0.346
	Lecture		0.779	0.236
	Sciences		0.807	0.174
	Résolution de problèmes		0.831	
	Générale	0.919		0.886
Suède	Mathématiques		0.894	0.395
	Lecture		0.771	0.252
	Sciences		0.805	0.233
	Résolution de problèmes		0.815	
	Générale	0.926		0.888