Chapitre 1

Évaluations à large échelle : prendre la juste mesure des effets de contexte

Dominique LAFONTAINE

1. INTRODUCTION

En posant que les enquêtes internationales à large échelle (ci-après ELA) ont toute leur pertinence, à condition de tenir compte des effets de contexte, ce chapitre revisite un vaste corpus d'enquêtes à large échelle, y situe les principales **avancées** et **points de rupture** tant sur le plan conceptuel que méthodologique qui font en sorte que les effets de contexte sont aujourd'hui, davantage et mieux qu'hier, pris en compte dans les enquêtes comparatives à large échelle. Si la question des effets de contexte est abordée au travers du prisme des enquêtes internationales, elle ne s'y réduit toutefois pas. Plusieurs des effets de contexte décrits dans le présent chapitre, en particulier dans les sections 2 et 3, s'observent et sont à prendre en considération dans des études à plus petite échelle, voire dans des approches plus qualitatives.

2. LES ÉVALUATIONS À LARGE ÉCHELLE

Le choix des évaluations comparatives à large échelle pour étudier les effets de contexte est facile à comprendre. Celles-ci sont en effet un terrain privilégié pour étudier la prise en compte et la mesure des effets de contexte pertinents ou indésirables (biais). On peut même soutenir que sans appréhension rigoureuse des effets de contexte, les enquêtes internationales sont sans intérêt, sans pertinence, voire risquent d'induire en erreur, si leur interprétation outrepasse les limites que leur imposent précisément les contextes. Un des reproches souvent émis à l'encontre des enquêtes internationales est qu'elles « comparent des pommes et des poires ». À ce reproche relativement élémentaire, on peut opposer que la comparaison de pommes et de poires peut faire sens, pour autant qu'on s'intéresse à l'arbre, au verger, au sol, aux conditions climatiques, aux traitements phytosanitaires... qui permettent de produire des fruits divers, non formatés, dans le respect de l'environnement et des populations locales.

Dans les évaluations comparatives à large échelle, les élèves ou les individus dont on mesure les performances scolaires ou cognitives sont issus de langues, de cultures, d'horizons et de systèmes éducatifs divers (diversité de contextes). Par ailleurs, ces individus sont nichés dans des classes, des filières, des programmes d'études, des écoles, des régions, des pays (contextes emboîtés). Ces deux types de contextes doivent être pris en compte de manière adéquate, tant sur le plan conceptuel que méthodologique.

2.1. Principales études internationales à large échelle

Lorsqu'on parle d'enquêtes internationales, on entend par là les grandes enquêtes comparatives menées principalement à l'initiative de l'Association internationale pour l'évaluation du rendement scolaire (IEA, voir http://www.iea.nl) et de l'Organisation de coopération et de développement économiques (OCDE, voir http://www.oecd.org). Le tableau suivant fournit un aperçu des principales enquêtes internationales menées depuis 1960. Dans ce tableau, on peut voir qu'une majorité de ces enquêtes portent sur les acquis ou le rendement des élèves dans les principaux domaines scolaires que sont la compréhension en lecture, les mathématiques ou les sciences. Toutefois, il arrive que d'autres domaines soient investigués, tels que l'éducation civique (Cived, ICCS), les technologies de l'information (Sites, ICILS) ou encore les enseignants (Talis). Par ailleurs, si les enquêtes internationales sont aujourd'hui surtout connues *via* la fameuse enquête PISA de l'OCDE, le tableau fait clairement apparaître que les enquêtes internationales se sont déployées dès la fin des années 1960, soit près de

trente ans avant que l'OCDE prenne des initiatives en la matière, et que même si l'IEA fait l'objet de moins d'attention médiatique, elle continue à jouer un rôle important dans le champ, notamment en évaluant d'autres populations cibles ou d'autres domaines que ne le fait PISA.

Tableau 1.1. Aperçu des principales enquêtes internationales menées depuis 1960

	1960	1970	1980	1990	2000	2010
Langue maternelle		Reading, Literature	Composition	RLS IALS (1998)	PIRLS 2001, 2006 PISA 2000, 2003, 2006, 2009 ALL 2003, 2006	PIRLS 2011, 2016 PISA 2012, 15, 18 PIACC 2010
Mathématiques	FIMS		SIMS IAEP I	IEAP II TIMSS, TIMSS_R	PISA 2000,03, 06,09 TIMSS 2003, 07 TIMSS Ad. 08	TIMSS 2011, 15 PISA 2012, 15, 18
Sciences		FISS	SISS IAEP I	IEAP II TIMSS, TIMSS_R	PISA 2000, 03, 06, 09 TIMSS 03, 07 TIMSS Ad. 08	TIMSS 2011, 15 PISA 2012, 15, 18
Autres		English, French as foreign ; Civic	Comped Classroom environment study PPP	CivED Sites-M1	Sites-M2 Teds-M (2008) ICCS 2009 Talis 2008 ESLC 2012	Talis 2013, 18 ICCS ICILS 13, 18

2.2. Prendre la juste mesure des effets de contexte

Le présent chapitre s'organise autour de la question d'une prise de mesure, considérée comme *juste*, des effets de contexte. Par « juste », il faut ici comprendre que cette prise en compte répond à une exigence de rigueur tant sur le plan conceptuel que méthodologique et que les réponses à cet impératif doivent être proportionnées et adéquates, compte tenu de l'état actuel des connaissances scientifiques. Cette prise en compte sera déclinée en trois phases : en amont des tests ou questionnaires, au moment du test et en aval du test. Ceci amènera à aborder des questions aussi diverses que les cadres de référence pour les questionnaires de contexte,

les notions d'équivalence (ou d'invariance[6]) et de stabilité cross-culturelle, les biais de méthodes/styles de réponse, ainsi que l'influence des groupes de référence et du macrocontexte « système éducatif ».

3. EN AMONT DES QUESTIONNAIRES

3.1. Cadres de référence conceptuels : penser les effets de contexte

L'analyse de l'évolution des Cadres de référence relatifs aux variables de contexte (*frameworks*) sera principalement menée au départ de l'enquête PISA. En effet, les politiques internes de gestion des enquêtes propres à l'IEA et à l'OCDE diffèrent sur plus d'un point. Sans entrer dans le détail de ces politiques, on peut souligner que l'OCDE agit de manière plus centralisée et directive, voire plus politique, au sens propre du terme ; il est donc plus aisé et plus pertinent de tenter une analyse des politiques de l'OCDE que de l'IEA. De son côté et par son statut, l'IEA est d'abord un consortium de centres de recherche et laisse davantage d'autonomie et de marge de manœuvre aux consortiums en charge de la mise en œuvre des enquêtes menées sous leurs auspices. La diversité d'approche d'une étude à l'autre y est donc plus grande et tient souvent davantage à l'empreinte que lui donne le consortium responsable plutôt qu'à une politique orchestrée par l'IEA.

La définition des fondements de l'enquête PISA, tant conceptuels que méthodologiques débute à la fin des années 1990. Un consortium de centres de recherches dirigé par l'*Australian Council of Educational Research* (ACER) se voit, après appel d'offre, confier la responsabilité de mettre en œuvre le programme PISA. Des groupes d'experts internationaux sont mis au travail et l'OCDE publie en 1999 son premier cadre de référence, intitulé *Mesurer les connaissances et compétences des élèves. Un nouveau cadre d'évaluation* (OCDE, 1999). Dans ce document fondateur, la focale est mise sur la mesure cognitive (savoirs et compétences) : une seule page (sur 85) est consacrée aux questionnaires de contexte. Certes, l'importance « d'indicateurs **contextuels** mettant en relation les performances avec les caractéristiques des élèves et des écoles » (OCDE, 1999, p. 10) est soulignée, mais aucun cadre de référence pour les questionnaires de contexte n'est produit ou rendu public. Aucune mention n'est faite du travail d'un groupe d'experts réfléchissant aux questionnaires de contexte, alors que ce groupe existe. Il travaille en quelque sorte dans l'ombre, ce qui a portée de symbole. Comme l'indique le rapport technique de *PISA 2000*,

6. Dans la littérature, on parle indifféremment d'équivalence ou d'invariance. Ici, nous utilisons le terme « équivalence » qui nous parait plus clair.

« le cadre de référence pour les questionnaires n'a pas été publié par l'OCDE, mais est disponible sous forme de document de travail » (p. 34)[7].

Pour les cycles 2003 et 2006, aucune évolution majeure n'est observée. Le QEG (*questionnaire expert group*) continue à travailler de manière relativement confidentielle, tandis que le travail des SMEG (*subject matter expert group*) est reconnu et mis en lumière dans les cadres de référence successifs (OCDE, 2003, 2006). Le point de rupture se produit lors du cycle PISA 2009. Un chapitre intitulé « Questionnaire framework », d'une longueur de 20 pages, apparaît dans le Cadre de référence de PISA 2009 (OCDE, 2009). Dans ce chapitre se déploie une conceptualisation systématique des constructs essentiels à prendre en considération au niveau système, école, classe et élève. En annexe, une liste exhaustive des constructs et des questions/items les mesurant dans les questionnaires est fournie.

Le fait que l'OCDE ait, pour PISA 2009, confié la tâche de développer les questionnaires de contexte au CITO[8], consortium distinct d'ACER, n'est sans doute pas étranger à cela. L'existence d'un groupe d'experts pour les questionnaires (QEG) dont le chairman est Jaap Scheerens (Université de Twente) est officialisée à la même époque. Sans surprise, le modèle conceptuel développé dans le Cadre de référence fait la part belle aux variables liées à l'école et au courant de la *School effectiveness research*, dont Jaap Scheerens est un illustre représentant. C'est aussi à la même époque que l'OCDE se met à publier, à côté du principal rapport présentant les résultats de PISA, des rapports approfondis consacrés à des questions de politiques éducatives – efficacité des écoles, équité, efficience… – qui nécessitent un modèle conceptuel plus strict des variables conceptuelles.

L'évolution entamée en 2009 se poursuit en 2012. Plusieurs évolutions retiennent l'intérêt lors de ce cycle. Eckhard Klieme (DIPF[9]) préside désormais le QEG. On note une systématisation du cadre conceptuel au départ duquel sont pensées d'une part la constitution d'une base de données à long terme (*sustainable database*), et la sélection d'un noyau de variables contextuelles (*core content*) pensé dans la durée (OCDE, 2012). Le cadre conceptuel, organisé selon le modèle CIPO (*context-input-processes-outcomes*), s'ouvre à des variables liées à la qualité de l'enseignement et des environnements d'apprentissage (opportunités d'apprentissage, feedbacks, gestion de la classe), questions dont Eckhard Klieme est un expert reconnu. L'apparition de variables contextuelles liées à la classe n'est pas en

7. « The questionnaire framework was not published by the OECD but is available as a project working document » (*Technical report PISA 2000*, p. 34).
8. Le CITO (Central Institute for Test Development) est une importance compagnie de testing, située à Arnhem aux Pays-Bas (http://www.cito.com).
9. Le DIPF (Deutsches Institut für Internationale Pädagogische Forschung), branche de l'Institut Leibniz, est un important centre de recherches situé à Francfort.

soi une innovation dans le cadre des enquêtes internationales. Les premières enquêtes de l'IEA dans les années 1970 ont même été un terreau fertile pour le développement de la notion d'OTL et du modèle de Carroll. Cependant, dans l'enquête PISA, qui rappelons-le, a fait le choix d'une population d'âge (élèves de 15 ans) et ne sélectionne pas de classes entières, mais des élèves à l'intérieur des écoles, il s'agit d'une évolution notable, témoignant du fait que PISA se préoccupe beaucoup plus qu'annoncé ou que prévu d'évaluer le rendement de l'école et l'impact des facteurs scolaires sur ce rendement.

Dans ce même chapitre du Cadre de référence de 2012 (OCDE, 2012), ainsi que dans le Rapport technique de PISA 2012 (OCDE, 2014), on voit pour la première fois apparaître une préoccupation pour la validité et la stabilité cross-culturelle des variables contextuelles. Dans ce cas, il s'agit d'une véritable avancée, à laquelle la présence dans le QEG de Fons van de Vijver (Université d'Utrecht), expert des questions de psychologie cross-culturelle, n'est sans doute pas étrangère. La question de l'équivalence et de la validité des comparaisons est certes centrale dans les enquêtes internationales, mais jusqu'alors, le questionnement et le débat, parfois vif et polémique (Goldstein, 2004 ; Lafontaine & Demeuse, 2002 ; Romainville, 2002), portait principalement sur l'équivalence des tests (mesures cognitives). Désormais, il s'agit de s'interroger sur la stabilité et l'équivalence des variables contextuelles : lorsque l'on mesure par exemple le concept de soi en mathématiques ou la motivation pour la lecture, mesure-t-on bien la même chose dans les différents contextes, et peut-on comparer de manière valide les scores des pays sur cette échelle ? Oser poser cette question dans le cadre des enquêtes internationales constitue un véritable point de rupture.

PISA 2015 voit se poursuivre les évolutions entamées en 2009 et 2012. L'OCDE elle-même consacre l'importance des questionnaires de contexte en lançant un appel d'offres séparé pour les domaines cognitifs (tests en lecture, mathématiques et sciences) et non cognitifs (questionnaires de contexte). Le DIPF emporte ce dernier marché ; Eckhard Klieme préside à nouveau le QEG. Pour la première fois, un cadre de référence propre est élaboré pour les questionnaires de contexte dont le titre lui-même est un programme : *The PISA 2015 Framework for Context Assessment. Monitoring Opportunities and Outcomes, Policies and Practices. Modelling Patterns and Relations, Impacts and Trends in Education*[10]. La boucle est en quelque sorte bouclée.

De 2000 à 2015, PISA a ainsi connu une montée en puissance des questionnaires de contexte, qui passent d'un statut quasi confidentiel à

10. Ce cadre de référence n'est pas encore publié sous ce titre par l'OCDE, 2012. Seul un *Draft Questionnaire framework* est accessible en ligne sur la page http://www.oecd.org/pisa/pisaproducts/pisa2015draftframeworks.htm.

la pleine consécration. Cette montée en puissance se traduit d'une part par une élaboration de cadres conceptuels de plus en plus rigoureux et étoffés, d'autre part par un souci croissant pour la qualité, la validité et la comparabilité cross-culturelle des variables contextuelles ou non cognitives (cycles 2012 et 2015). L'évolution des évaluations internationales et de PISA en particulier est en quelque sorte animée par un double mouvement. Ainsi, le focus se déplace progressivement des mesures cognitives (les tests eux-mêmes et le classement des pays qui en résulte) vers des résultats plus contextualisés qui tentent d'expliquer l'efficacité et l'équité des systèmes et les performances des élèves à la lumière des contextes[11]. Cette importance reconnue aux indicateurs contextuels s'accompagne d'une part de la nécessité de penser, à l'aide de cadres conceptuels aux assises théoriques solides, les variables contextuelles les plus importantes à investiguer, d'autre part de s'interroger sur la validité et la comparabilité de ces mesures contextuelles. Alors que la question brûlante de l'équivalence et de la comparabilité se limitait jusqu'en 2010 environ aux mesures cognitives (tests), celle-ci concerne désormais aussi les variables de contexte. Même s'il s'agit d'une bonne chose en termes d'avancée scientifique, ceci place assurément les enquêtes internationales devant des défis évaluatifs d'une importance jusque-là largement insoupçonnée.

4. DANS LES QUESTIONNAIRES

Dans cette section seront développés différents effets de contexte dont il est maintenant connu qu'ils peuvent affecter la validité et la comparabilité de certaines variables de contexte. On y traitera notamment des **biais de méthode** et **styles de réponse**, ainsi que de la notion d'**équivalence** et de **stabilité cross-culturelle**. Ces notions sont d'abord définies d'un point de vue théorique. On s'attache ensuite à décrire, à l'aide de données empiriques, comment ces biais de méthode et cette instabilité cross-culturelle peuvent être détectés lors des étapes de validation, prévenus lors de la conception des questionnaires, ou pris en compte *a posteriori* à l'aide de modélisations statistiques appropriées.

4.1. Biais de méthode et styles de réponse

Dans toute évaluation ou collecte de données, y compris qualitative, un manque de comparabilité ou une forme de « biais » peut résulter de la « **méthode** » utilisée, le terme méthode devant être ici entendu dans

11. En quelque sorte, PISA retrouve l'ambition initiale de Torsten Húsen qui, fondateur de l'IEA, voulait prendre le monde comme laboratoire pour comprendre le fonctionnement des systèmes éducatifs.

un sens très large, par ex. l'échantillonnage, le mode d'administration, ou l'instrument lui-même (pour une synthèse, voir Podsakoff, MacKenzie, Lee, & Podsakoff, 2003 ; van de Vijver & Tanzer, 2004). Pensons simplement aux examens tels qu'ils se pratiquent habituellement dans les universités. Si les étudiants ne sont évalués que par une seule « méthode », par exemple des examens écrits composés de questions à choix multiple, il en résulte une coloration ou un biais, voire une injustice dans l'évaluation des savoirs et compétences des étudiants. Certains en effet sont beaucoup plus à l'aise avec cette méthode que d'autres et le choix exclusif de cette modalité avantage donc systématiquement les mêmes étudiants. Par ailleurs, la mesure qui résulte des examens utilisant cette méthode sera différente de celle que donnerait un examen écrit avec réponses construites, un portfolio ou un examen oral.

Il en va de même dans les enquêtes internationales. La situation s'y complique du fait que la tendance à réagir à une « méthode » peut être différente selon les pays ou les cultures concernés. Comme signalé antérieurement, c'est à partir de 2009 que l'on trouve des traces dans les rapports des enquêtes internationales d'une préoccupation pour les biais systématiques liés aux styles de réponse engendrés en particulier par les mesures autorapportées mesurées par des échelles de Likert. Les recherches en psychologie cross-culturelle qui ont documenté ces phénomènes sont quant à elles bien antérieures (début des années 1990).

Les biais ou styles de réponse peuvent être définis comme « des tendances systématiques et stables dans la manière de répondre qui ne s'expliquent pas par le contenu d'une question ni par ce que celle-ci vise à mesurer » (Yang, Harkness, Chin & Villar, 2010, p. 203)[12]. Les principaux styles de réponse sont :

- l'acquiescement – tendance à « acquiescer avec les items quel que soit leur contenu » (Van de Vijver & He, 2014, p. 7) ou désacquiescement (tendance à s'opposer) ;
- le choix des extrêmes : tendance à utiliser de préférence les points extrêmes de l'échelle ;
- le choix des échelons intermédiaires : tendance à utiliser les échelons du milieu de l'échelle ou le point central s'il y en a un ;
- un comportement erratique, choix des réponses au hasard (par exemple, les choix correspondent à une forme telle qu'un escalier) ;
- la désirabilité sociale : tendance à adopter les réponses en conformité avec les normes ou les attentes sociales supposées.

12. Response styles are defined as « consistent and stable tendencies in response behavior that are not explainable in terms of question content or what a given question aims to measure » (Yang, Harkness, Chin & Villar, 2010).

Plusieurs de ces styles de réponse peuvent s'expliquer par un comportement peu engagé du répondant, dit « satisuffisant » (Krosnick, 1991) : celui-ci minimise son investissement cognitif en répondant d'une manière uniforme ou stéréotypée aux différents items d'une échelle ou même à plusieurs échelles d'un même questionnaire, faisant ainsi l'économie d'un traitement approfondi du contenu des différents items. Au-delà du comportement satisuffisant, certains comportements stéréotypés des répondants peuvent résulter d'un manque de maîtrise de la langue du questionnaire, ou de la charge cognitive que représente la lecture de certaines questions.

Nombre d'études ont établi que les styles de réponse varient en fonction de facteurs individuels (sexe et traits de personnalité) et de facteurs liés au pays ou à la culture. Ainsi, la tendance à l'acquiescement semble plus élevée aux USA (Harzing, 2006) et plus présente dans le sud que dans le nord de l'Europe (Harzing, 2006 ; van Herk, Poortinga, & Verhallen, 2004). La tendance à utiliser les extrêmes est plus élevée en Amérique du Nord que dans les pays asiatiques (Chen, Lee, & Stevenson, 1995 ; Heine *et al.*). Cette propension est parfois mise en relation avec la distinction entre une culture occidentale individualiste dans laquelle il est de bon ton d'avoir un avis individuel tranché, et une culture confucéenne/collectiviste où la norme est de se fondre dans la masse (Harzing, 2006 ; Hofstede, 2001 ; van Herk, Poortinga, & Verhallen, 2004 ; Yang, Harkness, Chin, & Villar, 2010).

Dans une enquête internationale, ceci constitue à l'évidence une menace pour la validité : la mesure du construct ciblé (par ex. : le concept de soi) est entachée par une tendance à réagir à l'instrument (échelle de Likert) variable selon les sujets répondants, mais aussi les cultures ou les pays. Il peut en résulter une certaine absence d'équivalence et de comparabilité des réponses.

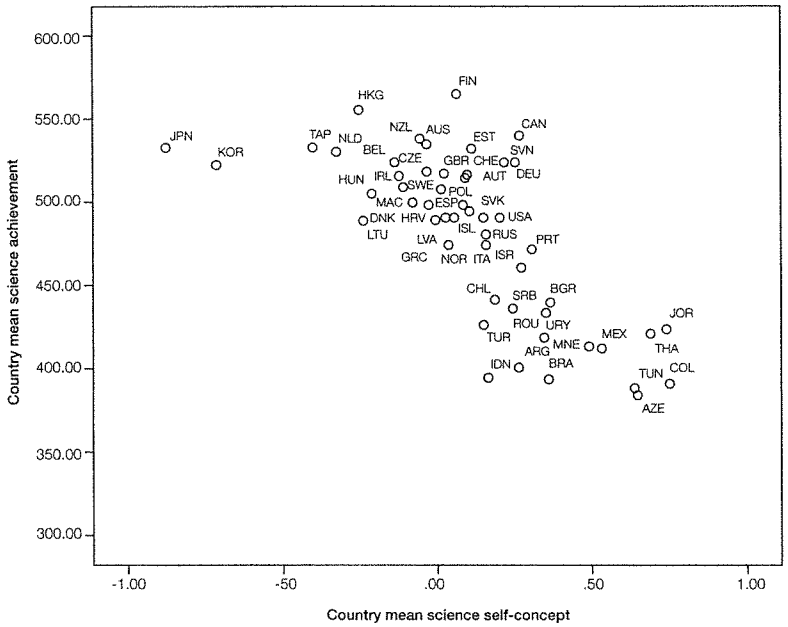
4.2. Le paradoxe attitudes-performances

À partir de 2005, plusieurs analyses secondaires menées sur les enquêtes internationales et PISA en particulier ont mis en évidence des biais de « méthode » pour les mesures autorapportées (échelles de Likert) : Buckley (2009) ; Grisay (2009 ; 2010) ; Kyllonen & Bertling, 2013, Lie & Turmo (2005), Schulz (2005) ; Van de Gaer, Grisay, Schulz, Gebhardt (2012). Ainsi, sur des échelles de Likert telles que celles mesurant l'intérêt pour la lecture ou le concept de soi en mathématiques ou en sciences, les biais de réponse (désirabilité sociale, tendance à choisir les extrêmes ou les intermédiaires, comportements « satisuffisants ») sont plus marqués dans certains pays.

Un phénomène maintes fois constaté agit comme un révélateur de l'existence d'un biais de méthode lié au pays : le « paradoxe Attitudes-Performances » (*attitudes-achievement paradox*). Pour cer-

tains constructs mesurant des « attitudes » (en fait, plus largement les mesures autorapportées où l'individu déclare sa motivation, son intérêt, son anxiété...), on s'attend *a priori* à ce que la corrélation entre les scores sur les échelles de ce type de construct soit de même signe (positive ou négative) que l'on calcule la corrélation à l'intérieur d'un pays ou qu'on la calcule au niveau des pays en corrélant dans ce cas les scores moyens des pays pour l'échelle d'attitudes avec les scores moyens des pays au test. Par exemple, on s'attend à une corrélation positive entre le concept de soi en sciences et les résultats au test de sciences : meilleurs sont les élèves au test, plus ils se perçoivent comme bons en sciences. Or, dans de nombreux cas, un tout autre résultat est observé, comme l'illustre le graphique ci-dessous.

Figure 1.1 Corrélation entre la moyenne par pays des résultats en sciences et du concept de soi en sciences dans PISA 2006 (Van de Gaer *et al.*, 2012)



Alors que la corrélation entre les attitudes et les performances est positive dans tous les pays, elle devient négative quand on calcule les corrélations au niveau pays (voir figure). Ce qui est pour le moins paradoxal, pour ne pas dire incompréhensible sauf à suspecter un biais systématique.

Cette corrélation négative au niveau pays signifie que c'est dans les pays les moins performants en sciences (dans le cas illustré, Colombie, Thaïlande, Jordanie, Tunisie, Kyrgyzstan, Azerbaidjan...) que les élèves déclarent les attitudes les plus positives envers les sciences. À l'inverse, parmi les pays déclarant un concept de soi en sciences particulièrement bas, on retrouve le Japon, la Corée, Taipei, Hong Kong... qui affichent d'excellentes performances en sciences, mais sont aussi connus pour leur biais de modestie. Voilà qui n'a pas manqué d'intriguer les chercheurs et qui les a mis sur la piste des biais liés aux styles de réponse. Nous reviendrons plus longuement ci-après sur cette question lorsque seront envisagés les moyens de contrer les biais liés aux styles de réponse.

4.3. Équivalence cross-culturelle

Pour repérer si une échelle est équivalente d'un groupe à l'autre (qu'il s'agisse de pays ou d'autres groupes tels que les garçons *vs* les filles, les natifs *vs* les immigrés...), il existe différentes procédures, dont la plus connue est l'analyse factorielle confirmatoire multigroupes (MG-CFA : *multigroup confirmatory factorial analysis*)[13]. Cette procédure a tendance à se généraliser aujourd'hui dans les enquêtes internationales (on en trouve des traces dans les rapports techniques), ainsi que dans les études qui portent sur la validation d'un instrument.

Sans entrer dans les détails techniques, ces modèles testent trois types d'équivalence :

- l'équivalence **configurale** : une même structure factorielle doit être observée dans les différents pays ; ceci garantit qu'un même construit est mesuré dans les différents pays ;
- l'équivalence **métrique** : « l'unité de mesure doit être identique pour les échelles, même si l'origine est différente (...)[14] » (Arends-Toth,

13. « Dans les études CFA avec des groupes multiples, il est possible de tester l'invariance d'un ensemble ou de toutes les estimations de paramètres à travers les différents groupes. Les tests de l'invariance factorielle (...) définissent traditionnellement une série de modèles nichés dans lesquels les extrémités sont le modèle le moins restrictif, sans aucune contrainte d'invariance, et le modèle le plus restrictif (invariance totale) où tous les paramètres sont contraints à être identiques à travers tous les groupes. Tester pour l'invariance factorielle revient essentiellement à comparer un ensemble de modèles dans lesquels des aspects de la structure factorielle sont maintenus invariants à travers les groupes, et à évaluer les indices de fit quand les éléments de ces structures sont contraints » (Marsh, Tau, Artelt, Baumert, & Peschar, 2006, p. 331).

14. « Un exemple est la mesure de la température à l'aide des échelles de Celsius et de Kelvin. Les deux échelles mesurent le même concept et les unités de mesure sont identiques, mais les origines des échelles ne le sont pas (les degrés Kelvin peuvent être convertis en degrés Celsius en ajoutant 273 à la température mesurée en degrés Kelvin » (Arends-Toth, Van de Vijver, & Poortinga, 2006, p. 4).

Van de Vijver, & Poortinga, 2006, p. 4). Si ce niveau d'équivalence est atteint, le(s) facteur(s) peu (ven) t être corrélé(s) avec d'autres variables de façon valide, et il est légitime de comparer ces corrélations d'un pays à l'autre ;

- l'équivalence **scalaire** : « non seulement la métrique, mais un autre type de valeur standard, généralement l'origine de l'échelle, doit être identique » (Arends-Toth, Van de Vijver, & Poortinga, 2006, p. 4). Si ce niveau d'équivalence est atteint, les moyennes des pays sur le(s) facteur(s) peuvent être comparées d'un pays à l'autre. L'équivalence scalaire est très difficile à atteindre, et dans les faits, est rarement rencontrée[15].

Si on reprend l'exemple relatif au concept de soi en sciences, l'équivalence scalaire n'est pas atteinte dans ce cas, ceci signifie qu'il n'est pas valide de dire que les élèves japonais ont un concept de soi en sciences nettement moins bon que celui des élèves tunisiens. En revanche, l'équivalence métrique étant atteinte, on peut avancer que la corrélation entre le concept de soi et les performances en sciences est plus élevée au Japon qu'en Tunisie. Ceci, on l'aura compris, met des restrictions sérieuses sur le type de conclusions que l'on peut tirer de certaines échelles de mesure abondamment utilisées dans les enquêtes à large échelle, ou même à plus petite échelle.

4.4. Comment prendre en compte les biais de méthode ou neutraliser les effets de contexte indésirables ?

On distingue à cet égard deux grands types d'approches :

- dès la **conception du questionnaire** : on peut agir en privilégiant des formats d'items alternatifs tels que des scénarios ou des choix forcés, recourir à des vignettes d'ancrage (*anchoring vignettes*), ou utiliser des items « piégeant » les biais de réponse (« topic familiarity and foils » dans PISA 2012) ;
- **après-coup** : on recourt à différentes techniques de modélisation du biais de méthode/styles de réponse (calcul d'indices ERS, ARS, « superconstruct » = acquiescence, modélisations IRT) pour corriger les résultats observés en les « purifiant » du biais de réponse.

15. « It is uncommon to find support for scalar invariance in studies involving huge samples in many countries because the invariance tests are sensitive to sample size » (Van de Vijver & He, 2014, p. 17).

4.4.1. *Approche* a priori *: format d'items alternatifs et détection des styles de réponse*

Dès PISA 2012 et 2015, conscients des biais de méthodes liés aux échelles de Likert, les personnes en charge du développement des questionnaires ont exploré une série d'alternatives. Ainsi, pour certaines mesures autorapportées, à côté du format habituel (échelle de Likert), une échelle avec les mêmes items, mais forçant le choix entre deux stratégies, a été testée. La version avec choix forcé augmente légèrement la corrélation avec le rendement dans chaque pays, mais change radicalement la corrélation au niveau pays : la corrélation passe ainsi de -0.47 (avec l'échelle de Likert) à + 0.60 (version avec choix forcé) (Kyllonen & Bertling, 2013). Cette forte corrélation négative dans la version Likert est une nouvelle illustration du paradoxe attitudes-performances. La version avec choix forcé voit quant à elle disparaître le paradoxe.

Pour une autre échelle portant sur les opportunités d'apprentissage en mathématiques (PISA 2012), une liste de 19 termes mathématiques est fournie aux élèves (logarithme, probabilités, cosinus, aire...) dans laquelle quelques termes fictifs ont été inclus (échelle subjonctive, nombre propre, fraction déclarative...). Il est demandé aux élèves d'indiquer leur degré de familiarité avec chacun des concepts (de « jamais entendu » à « je connais et je comprends le concept »). Pour les items « piège », le degré de familiarité devrait être nul, vu que ces concepts n'existent tout simplement pas. Deux scores sont calculés pour l'échelle d'OTL : un score **brut** avec tous les concepts OTL réels ou fictifs et un score **ajusté** en soustrayant la moyenne pour les faux concepts de la moyenne pour tous les concepts. La comparaison des corrélations entre les deux scores, brut et ajusté, et les performances en mathématiques fait apparaître les phénomènes déjà décrits ci-dessus. À l'intérieur des pays, les corrélations entre les deux scores de familiarité et le score en maths dans chaque pays sont quasi identiques (la moyenne OCDE des corrélations par pays est respectivement de 0.45 et 0.44). En revanche, les corrélations au niveau pays se distinguent nettement. Alors que la corrélation entre le score brut d'OTL et le score en mathématiques n'est que de 0.17, avec le score ajusté elle est de 0.54. Si dans le présent cas, la corrélation entre les OTL (score brut) et les performances n'est pas négative, elle est néanmoins anormalement basse. Ceci tient une nouvelle fois au fait que dans certains pays plus que dans d'autres, les élèves ont tendance à indiquer plus souvent qu'ils connaissent des concepts mathématiques **inexistants**, par acquiescence, désirabilité sociale ou comportement satisfaisant, ou tout simplement parce qu'ils ne lisent pas les items.

Une troisième alternative utilisée dans PISA 2012 et 2015 (et dans nombre d'autres surveys) consiste à utiliser ce que l'on appelle des vignettes d'ancrage (*anchoring vignettes*). Cette technique consiste à

faire évaluer une série de vignettes dites d'ancrage et à s'en servir pour ajuster les scores des individus sur d'autres échelles qui concernent leur propre perception.

Ainsi dans PISA 2012, trois vignettes d'ancrage ont été créées en lien avec l'échelle relative au soutien des enseignants. Ces vignettes concernent des enseignants fictifs, qui n'ont rien à voir avec les enseignants des élèves ; celles-ci, à la limite de la caricature, tentent d'appréhender la propension des élèves à juger différemment ces enseignants au comportement typé. Il leur est demandé d'indiquer dans quelle mesure les trois enseignants ci-dessous se soucient des apprentissages de leurs élèves :

- *Mme Demonty donne des devoirs de maths presque tous les jours. Elle rend toujours les devoirs corrigés avant les examens. Mme Demonty se soucie des apprentissages de ses élèves.*
- *M. Marcoux donne des devoirs de maths une fois par semaine. Il rend toujours les devoirs corrigés avant les examens. M. Marcoux se soucie des apprentissages de ses élèves.*
- *Mme Vlassis donne des devoirs de maths une fois par semaine. Elle ne rend jamais les devoirs corrigés avant les examens. Mme Vlassis se soucie des apprentissages de ses élèves.*

Les vignettes d'ancrage servent à détecter la tendance des répondants à juger de manière modérée ou extrême les vignettes communes (profs hypothétiques) et à discriminer les items. Les trois vignettes d'ancrage servent ensuite pour ajuster les réponses des mêmes répondants à l'échelle relative au soutien de **leur** enseignant (***Mon*** *prof de maths fournit de l'aide supplémentaire si nécessaire...* : 4 items avec 4 degrés d'accord). Sachant ce que l'on sait des styles de réponse, il est plus que probable que dans certains pays, les élèves feront peu de différences entre les trois professeurs fictifs, tandis que d'autres pays, les élèves porteront des jugements plus contrastés.

Les corrélations entre le score sur l'échelle de soutien par l'enseignant et le score en maths ont été calculées pour 63 pays (Kyllonen & Bertling, 2013). Pour l'échelle de soutien, on dispose d'un score brut et d'un score ajusté grâce aux vignettes. À l'intérieur de chaque pays, sans surprise, les corrélations entre le score en maths et le score ajusté et non ajusté sont proches (0.03 et 0.13 respectivement). En revanche, au niveau des pays, la différence est impressionnante. La corrélation entre le score de soutien non ajusté et les performances en maths est de – 0.45, un spectaculaire paradoxe attitudes-rendement est observé : c'est dans les pays où les élèves se disent davantage soutenus par leur enseignant que les performances sont les plus faibles. Lorsque le score de soutien est ajusté grâce aux vignettes, pour tenir compte de la propension des élèves à donner des réponses extrêmes ou désirables, la corrélation « se redresse » et devient positive (0.29).

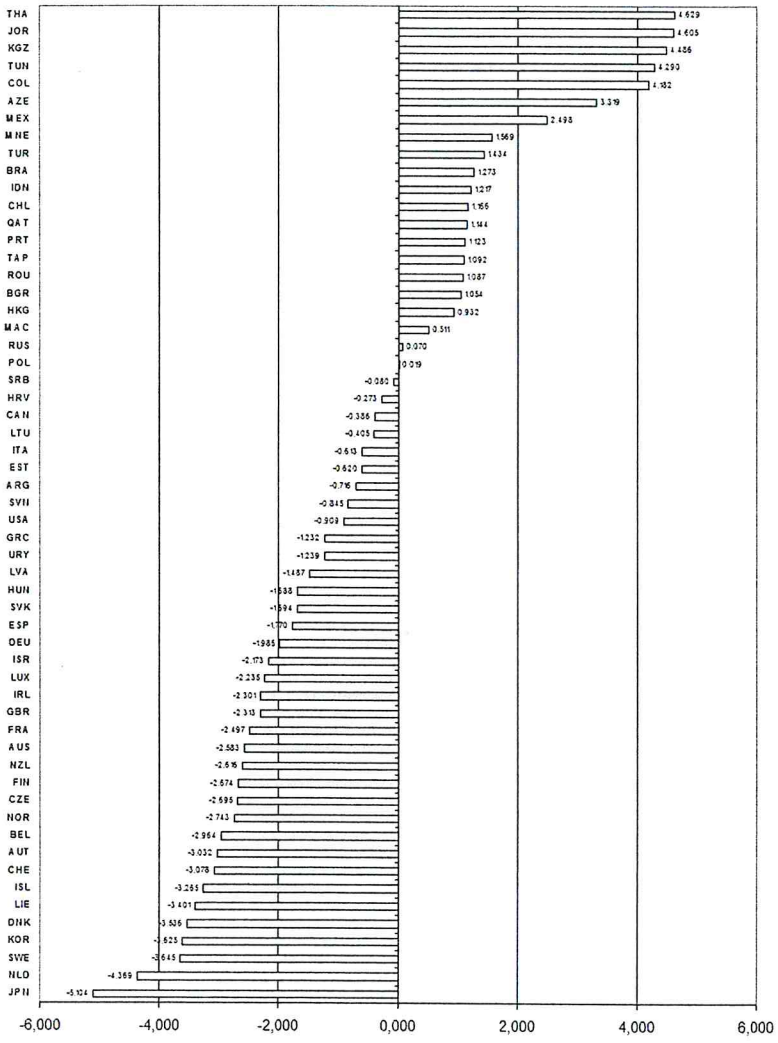
Les trois procédures décrites ci-avant montrent que si d'indéniables biais de réponses mettent du bruit dans les mesures autorapportées sur des échelles de Likert, des solutions alternatives existent qui permettent de prévenir les biais dès la conception du questionnaire. Cependant, dans un certain nombre de cas, il est difficile de recourir à d'autres formats que les échelles de Likert, ou il est trop tard, car les données ont déjà été collectées. On se tourne alors vers l'application de procédures *a posteriori*.

4.4.2. *Approche* a posteriori *: modélisation des biais*

Dans cette approche après coup, des modélisations statistiques sont appliquées pour tenter d'isoler la tendance des sujets, des groupes ou des pays à donner des réponses socialement désirables (Podsakoff *et al.*, 2003). Nous allons en donner deux exemples.

Lie & Turmo (2005) appliquent ainsi une analyse factorielle à 10 échelles de Likert dans PISA 2003 mesurant une série de constructs passablement différents (self-concept, relations élèves-enseignants, coopération et compétition, attitudes envers l'école...). De manière un peu surprenante, les analyses font apparaître des saturations élevées de tous les items sur un facteur général, commun à l'ensemble des échelles. Les auteurs considèrent que ce facteur principal, qu'ils appellent « superconstruct », correspond au style de réponse des élèves, à leur tendance à acquiescer, donner des réponses extrêmes, ou encore socialement désirables, quel que soit le contenu de la question. Des scores élevés du pays sur le superconstruct indiquent que les élèves dans ce pays ont en moyenne tendance à donner des réponses d'un certain style, quel que soit le construct mesuré. Le superconstruct explique 66 % de la variance des scores sur les 10 échelles.

Figure 1.2. Scores des pays sur le superconstruct de Lie et Turmo fondé sur 10 échelles d'attitudes (PISA 2003)



Tout en haut du superconstruct, avec des scores très élevés, on trouve les pays suivants : Thaïlande, Jordanie, Kyrgyzstan, Tunisie, Colombie, Azerbaidjan, Mexico, Montenegro, Turquie, Brésil, Indonésie, Chili, Quatar, Portugal, Taipei, Roumanie... Tout en bas du superconstruct figurent le Japon, les Pays-Bas, la Suède, la Corée, le Danemark, le Liechtenstein, la Suisse, l'Autriche, la Belgique, la Norvège, la Tchéquie, la Finlande, la Nouvelle-Zélande, l'Australie, la France... Dans le premier groupe, on retrouve une majorité de pays partenaires (hors OCDE), pays d'Amérique du Sud, du nord de l'Afrique, d'Asie de l'Est, d'Europe du Sud et par ailleurs assez peu performants au test PISA ; dans le second groupe, on retrouve les pays asiatiques de l'OCDE, les pays scandinaves, du centre, du nord ou de l'est de l'Europe, l'Australie et la Nouvelle-Zélande, pays plus performants au test PISA dans l'ensemble. La similitude des patterns avec la figure illustrant le paradoxe attitudes-performances est à relever, de même qu'avec la littérature sur les styles de réponse cross-culturels.

Lie et Turmo utilisent ensuite le superconstruct pour ajuster le score des différentes échelles concernées, appliquant en quelque sorte un coefficient modérateur aux répondants au style extrême et un coefficient amplificateur à ceux qui se montrent tièdes ou prudents, selon un principe proche de celui des vignettes d'ancrage. En utilisant le superconstruct pour ajuster ainsi le score des pays sur les échelles de Likert, les corrélations au niveau pays avec le rendement deviennent nettement moins négatives, voire positives. La corrélation entre le score et le concept de soi en maths passe ainsi de – 0.26 à + 0.74.

Lafontaine, Baye, Vieluf et Monseur (2015) ont appliqué un autre type de modélisation à une échelle de 17 items (échelle de Likert) mesurant les opportunités d'apprentissage (OTL) en lecture dans PISA 2009. Le but de leur étude était double : ils voulaient d'une part améliorer la validité de la mesure du construct visé (OTL en lecture), d'autre part mieux comprendre à quoi peuvent correspondre les biais ou styles de réponse dans PISA. Leur étude s'est déroulée en plusieurs étapes en conservant 11 items de l'échelle initiale.

Dans un premier temps, une analyse factorielle exploratoire a été réalisée. Deux facteurs principaux ont été extraits, un facteur « fiction » (les enseignants utilisent des textes de fiction et posent des questions à leur propos) et un facteur « textes non continus » (les enseignants utilisent comme support des textes contenant des tableaux, graphiques, schémas et posent des questions à leur propos). Les deux facteurs ont été corrélés avec le score en lecture au niveau individuel, école et pays.

Tableau 1.2. Corrélations des deux facteurs OTL avec les scores en lecture (PISA 2009)

	Fiction	Textes non continus
Élève	0.17	0.10
École	0.33	0.17
Pays	– 0.32	0.27

Les corrélations obtenues au niveau Élève et École ne sont guère élevées ; pour des variables telles que les OTL, on s'attendrait à ce qu'elles le soient davantage. Au niveau pays, le paradoxe attitudes-performances est observé pour le facteur Fiction ; la corrélation avec le score lecture est de -0.32. Au vu de ces résultats, un biais de méthode est suspecté. Les auteurs soumettent alors les données à une analyse IRT *within-item* tridimensionnelle[16]. Tous les items de l'échelle OTL sont alloués à la dimension 1. L'hypothèse, un peu comme dans l'étude de Lie et Turmo (2005), est que cette dimension capte ce qui est commun à tous les items d'OTL, dont un style de réponse général. Les deux autres dimensions correspondent quant à elles aux deux sous-dimensions du construct d'OTL (« fiction » et « textes non continus ») « épurées » du biais de réponse. Ces trois dimensions sont ensuite corrélées à leur tour avec le score de lecture.

Tableau 1.3. Corrélations entre les trois dimensions de l'analyse IRT et les scores en lecture (PISA 2009)

	Style de réponse	OTL fiction	OTL textes non continus
Élève	– 0.24	0.40	0.32
École	– 0.37	0.51	0.42
Pays	– 0.50	0.05	0.48

La dimension 1 (style de réponse) corrèle de manière négative et importante, surtout au niveau pays, avec les performances en lecture. C'est

16. Parmi les modèles IRT multidimensionnels, on distingue des modèles *between-item* et des modèles *within-item*. Dans les modèles *between-item*, chaque dimension rassemble des groups d'items distincts. Dans les modèles *within-item*, chaque item est relié à plus d'une dimension. Typiquement, tous les items d'une échelle sont alloués à une dimension générale, et certains des items sont en outre alloués à une ou plusieurs dimensions spécifiques, selon le nombre de dimensions du modèle.

donc dans les pays les moins performants que les élèves se déclarent le plus exposés à des OTL *a priori* favorables aux apprentissages. On nage à nouveau en plein paradoxe… Si la dimension 1 mesure bien, comme nous le supposons, le style de réponse, c'est donc dans les pays les moins performants que le biais est le plus marqué, ce qui explique le paradoxe. Par ailleurs, pour les dimensions d'OTL épurées du biais de réponse, les corrélations avec le score en lecture sont positives et plus robustes qu'avec les facteurs issus de l'analyse factorielle exploratoire, dans lesquels la dimension OTL était « polluée » par le style de réponse. En ce qui concerne les corrélations au niveau pays, la corrélation négative de la dimension « fiction » devient légèrement positive (0.05), et la corrélation de la dimension « textes non continus » avec la lecture passe de 0.27 à 0.48.

Pour valider dans quelle mesure la dimension 1 capte bien un style de réponse général, les auteurs ont ensuite calculé deux indices classiques de styles de réponse au départ d'un ensemble de 41 autres items du questionnaire PISA (échelles de Likert) mesurant différents constructs (cet ensemble de 41 items comprend 7 items inversés) :

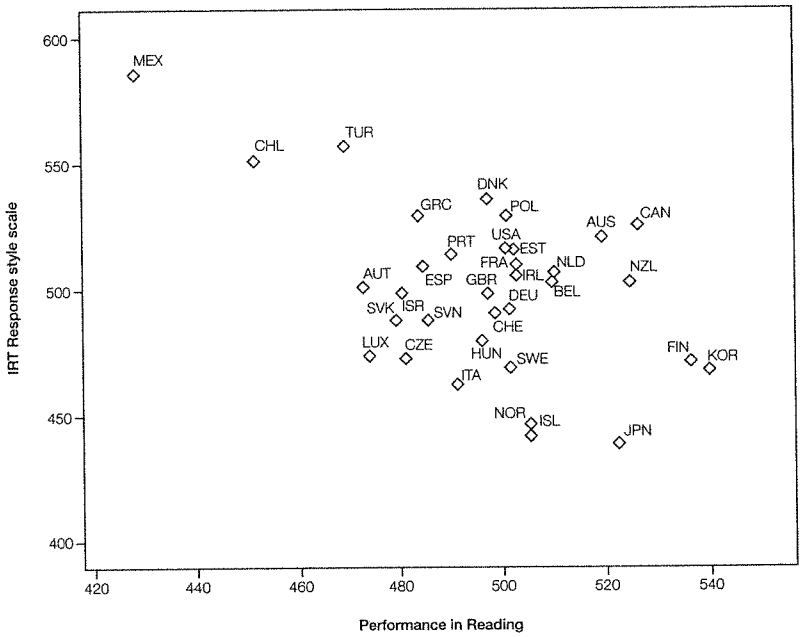
- un indice d'*acquiescence* (ARS) reprenant les % d'élèves qui se disent d'accord ou tout à fait d'accord avec les items, que ceux-ci soient formulés positivement ou négativement ; ceci est une attitude paradoxale, on ne peut pas à la fois se déclarer tout à fait d'accord avec un item comme « j'adore la lecture » et un autre qui dit « je déteste la lecture » ;
- un indice de *désirabilité sociale* (SDRS) reprenant les % d'élèves qui se disent d'accord ou tout à fait d'accord avec des items positifs et pas d'accord/pas du tout d'accord avec les items négatifs ; dans ce cas, l'élève se conforme, mais son attitude n'est pas absurde, il traite différemment les items selon leur contenu.

Sans surprise, les deux indices de style de réponse sont fortement corrélés avec la dimension 1 « style de réponse » (respectivement 0.62 et 0.55) et négativement liés aux performances en lecture (-0.48 et -0.42). La dimension 1 capte bien, dans une large mesure, un style de réponse général. Ceci est conforme aux données de la littérature. Selon Van de Vijver & He (2014), il existe en effet des preuves de l'existence d'un biais de réponse général.

Lafontaine *et al.* (2015) se sont ensuite penchés sur le lien entre les performances en lecture et le style de réponse au niveau individuel, école et pays. Une corrélation négative modérée à forte est observée aux trois niveaux entre le style/biais de réponse et les performances (corrélations de -0.24 au niveau individuel, -0.37 au niveau école et -0.50 au niveau pays). En clair, ceci signifie que ce sont les élèves les plus faibles au test de lecture,

les écoles les moins performantes et les pays les moins performants qui présentent le biais de réponse le plus marqué, comme des études rapportées ci-avant l'ont aussi montré (Lie & Turmo, 2005 ; Van de Gaer *et al.*, 2012).

Figure 1.3. Corrélation entre la moyenne par pays de la performance en lecture et la moyenne de la dimension style de réponse (PISA 2009, pays de l'OCDE uniquement)



Des patterns de pays sont à nouveau observés, même si cette analyse se limite aux pays de l'OCDE. On retrouve parmi les pays performants où le biais de réponse est peu marqué les pays scandinaves et les pays asiatiques. Parmi les pays peu performants où le biais de réponse est marqué figurent deux pays d'Amérique du Sud, la Turquie et la Grèce.

De ces deux études implémentant des modélisations statistiques *a posteriori*, on retiendra que des biais/styles de réponse affectent bien certaines des échelles d'une manière systématique et non aléatoire (patterns « culturels » et autres tels que le niveau de performances). Ces patterns épousent d'assez près ce que les recherches en psychologie culturelle ont établi (pour une synthèse, voir Yang, Harkness, Chin, & Villar, 2010) : les pays du sud (Amérique, Europe, Afrique) présentent en général des biais

de réponse plus marqués que les pays du nord, qui sont aussi souvent plus performants et plus riches. La bonne nouvelle est qu'il existe des techniques permettant de contrôler en partie ces biais résultant d'une interaction entre la méthode (échelles de Likert) et le contexte (pays, culture) et de se débarrasser ainsi de certains résultats paradoxaux.

5. EN AVAL DU TEST

Dans cette troisième et dernière section, nous aborderons la prise en compte des contextes telle qu'elle se pose une fois les données collectées, pour analyser et interpréter de manière rigoureuse la manière dont les micro-contextes (classe, école) et le macrocontexte (système éducatif) expliquent ou éclairent les phénomènes observés. Ces questions seront abordées au travers de deux études empiriques en lien avec les théories de la comparaison sociale.

Dans les enquêtes internationales, mais aussi dans la plupart des études portant sur les variables scolaires, il importe de tenir compte de l'emboîtement des contextes : les élèves sont nichés dans des classes, elles-mêmes nichées dans des écoles et les écoles s'inscrivent dans un système éducatif qui a ses caractéristiques propres. Pour estimer correctement les effets propres à chacun des niveaux, il importe de tenir compte de ces emboîtements dans les analyses statistiques, en appliquant des modèles de régression multiniveaux qui tiennent compte de cette hiérarchie (Bressoux, 2013)[17]. Par ailleurs, cet emboîtement a aussi un impact sur les comparaisons et les cadres de référence qu'un élève mobilise lorsqu'il s'autoévalue. Si on demande à un élève s'il est bon en mathématiques, celui-ci, pour répondre, en réfère inévitablement aux élèves qui l'entourent. C'est à ce type de phénomène que nous allons maintenant nous intéresser.

En psychologie sociale, les théories de la comparaison sociale ont abondamment montré que l'auto-évaluation dépend du contexte et des groupes dont on fait partie. Marsh et ses collaborateurs ont publié de nombreux travaux portant sur le concept de soi **académique** (ASC « *academic self-concept* ») mesuré à l'aide d'items tels que *je suis fort en maths, en sciences...* Ceci suppose une comparaison au moins implicite par rapport à d'autres élèves de la classe ou de l'école que l'on fréquente. Ces travaux ont contribué à populariser le BFLPE – « *big-fish-little-pond-effect* » : le poisson qui nage dans un petit étang se sent plus gros que s'il nageait dans l'océan. En étudiant différents contextes, et en utilisant notamment les données

17. Ce point ne sera pas développé dans le cadre de ce chapitre, mais l'apparition des modélisations multiniveaux dans les années 2000 compte au rang des évolutions notables pour mesurer de manière plus adéquate qu'auparavant les effets de contexte.

PISA (Marsh & Hau, 2003 ; Seaton, Marsh, & Craven, 2010), Marsh et ses collaborateurs ont pu montrer qu'au niveau individuel, le concept de soi académique est évidemment corrélé positivement avec les performances : un élève plus performant tend à avoir un meilleur concept de soi en mathématiques. Toutefois, au niveau école, à compétences égales, on observe un effet négatif de la moyenne des performances de l'école sur le concept de soi. Autrement dit, si on compare des élèves qui obtiennent des performances similaires au test PISA, mais dont certains fréquentent une école plus performante et d'autres une école moins performante, ceux qui fréquentent l'école moins performante ont un concept de soi **plus** favorable que ceux fréquentant une école plus performante. Selon Marsh, cet effet BFLPE est robuste, généralisé et universel. Il concerne aussi d'autres variables que le concept de soi comme les aspirations ou le choix d'études (Marsh & O'Hara, 2010 ; Nagengast & Marsh, 2012). Dans des travaux plus récents menés sur des données néerlandaises, Marsh, Kuyper, Morin, Parker, & Seaton (2014) ont montré que c'est surtout la classe fréquentée, davantage que l'école, qui influence le concept de soi, en tout cas dans un système différencié comme celui des Pays-Bas. Trautwein, Lüdtke, Marsh, & Nagy (2009) ont en outre montré, en travaillant sur un échantillon de données allemandes, que ce sont les élèves les plus faibles qui sont les plus sensibles aux effets négatifs sur le concept de soi qu'engendre le fait d'être dans un groupe-classe performant.

Cet effet ne concerne pas que les études internationales, il est présent dans tout type d'évaluation, même interne et locale, dès lors qu'un individu doit s'autoévaluer ou rendre compte de ses pratiques, et que pour ce faire, il utilise potentiellement un groupe de référence.

Pensons à quelque chose d'aussi simple que l'auto-évaluation de sa taille : comment chacun juge-t-il sa taille sur une échelle du type « je suis très petit/plutôt petit/plutôt grand/très grand » ? À taille identique, l'auto-évaluation risque d'être sensiblement différente selon que l'on est un homme ou une femme, Hollandais ou Vietnamien, issu d'une famille de grands ou de petits…

Examinons maintenant au travers de deux études empiriques comment ces cadres de référence jouent dans les enquêtes internationales et si le BFLPE est aussi universel que Marsh le soutient. Lafontaine & Monseur (2007) se sont penchés sur la manière dont un certain nombre de variables motivationnelles liées aux mathématiques (concept de soi, motivation extrinsèque et intrinsèque, sentiment d'auto-efficacité, anxiété) étaient corrélées avec les scores en mathématiques dans PISA 2003. Pour le concept de soi (évaluation de ses capacités en mathématiques, « je suis bon en mathématiques), par exemple, la corrélation moyenne dans les pays de l'OCDE est de 0.36 en moyenne, mais une grande variabilité de cette corrélation est observée selon les pays. Ainsi, la corrélation est seulement de 0.22 en

Belgique tandis qu'en Norvège, elle atteint 0.56. L'ampleur de cette variation pose question. Il apparaît en outre que cette variation n'est pas aléatoire : les corrélations sont en effet plus élevées dans les systèmes éducatifs dits compréhensifs ou non différenciés, à savoir les systèmes dans lesquels les élèves ne sont pas regroupés en filières à l'âge de 15 ans, et où le regroupement d'élèves dans des classes ou des écoles dont le niveau d'aptitudes est très variable est nettement moins prégnant que dans les systèmes dits différenciés ou stratifiés, qui pratiquent le regroupement des élèves par niveau. La variation semble donc liée aux caractéristiques des systèmes éducatifs, et donc au contexte.

Plus précisément, Lafontaine & Monseur (2007) font l'hypothèse d'un double effet de contexte :

- les élèves, lorsqu'ils s'autoévaluent, utilisent un point de comparaison – leur groupe classe : un élève moyen se sentira plus fort dans une classe faible que dans une classe très forte (BFLPE) ;
- la composition des groupes classe/école étant elle-même plus ou moins variable (homogène/hétérogène) selon la structure du système éducatif, l'effet de contexte 1 (lié à la classe ou à l'école) se traduit différemment selon le contexte 2 (système éducatif).

Dans les systèmes éducatifs compréhensifs (pas de filières, peu de redoublements, variance entre écoles faible), les groupes de référence classe et école servant de base de comparaison diffèrent peu entre eux. En revanche, dans les systèmes différenciés, par définition, les groupes de référence ont des profils académiques très contrastés (écoles/classes très, moyennement ou très peu performantes). Le point de comparaison (norme) que prennent les élèves pour s'autoévaluer y est donc éminemment variable, et les effets liés au groupe de référence y sont en conséquence démultipliés.

Pour tester leurs hypothèses, les chercheurs ont standardisé les variables motivationnelles par école/filière/classe d'étude (variables centrées et réduites). La standardisation neutralise l'impact du côté variable du groupe de référence ; après standardisation, les scores des élèves sur les échelles de motivation sont définis en termes d'écarts à la moyenne de leur groupe plutôt que sous forme de valeurs absolues. Les corrélations avant/après standardisation entre les scores sur les échelles de motivation et le score en mathématiques ont ensuite été comparées. Avant d'entamer la présentation des résultats, signalons que la comparaison à un groupe de référence joue potentiellement pour toutes les variables motivationnelles, même si c'est de façon implicite, sauf pour une, le sentiment d'auto-efficacité. À la différence du concept de soi (*je suis bon ou meilleur que mes camarades en maths*), la question porte sur l'auto-évaluation par l'élève de sa capacité à effectuer des tâches de mathématiques précises

(par ex. : je suis capable de calculer une ristourne de 15 % sur un article). Dans ce cas, l'élève n'a aucune raison de s'en référer à un groupe pour juger de sa capacité à effectuer ce type d'opération.

Tableau 1.4. Corrélations entre les variables motivationnelles et le score en mathématiques avant/après standardisation par école/filière/année d'étude (PISA 2003)

		Motivation intrinsèque	Motivation extrinsèque	Concept de soi	Anxiété	Auto-efficacité
Norvège	***Brute***	0.40	0.32	0.56	– 0.50	0.55
	Standard	0.40	0.30	0.56	– 0.49	0.55
Belgique	***Brute***	0.14	0.11	0.22	– 0.24	0.43
	Standard	0.21	0.17	0.35	– 0.30	0.36
Médian OCDE	***Brute***	0.14	0.12	0.35	– 0.34	0.51
	Standard	0.21	0.17	0.37	– 0.33	0.40

Pour des raisons de lisibilité ne sont repris dans le tableau qu'un système éducatif non différencié (la Norvège) et un système éducatif différencié (la Belgique), ainsi que le médian OCDE. Le pattern de résultats qui se dégage est très clair : standardiser les variables ou pas en Norvège ne change rien. Comme les différences de performance entre groupes (école, filière) y sont peu marquées, le point de comparaison y est déjà assez semblable d'une école à l'autre, avant standardisation. En Belgique, en revanche, les corrélations après standardisation sont sensiblement plus élevées que les corrélations brutes, sauf précisément pour le sentiment d'auto-efficacité. L'augmentation de la corrélation est plus nette pour le concept de soi qui, de toutes les variables motivationnelles, est celle où la référence au groupe est la plus explicite (je suis meilleur en maths que la plupart des élèves de ma classe).

La différence de résultats entre les variables motivationnelles, en particulier le concept de soi d'une part et le sentiment d'efficacité d'autre part, est la preuve que les élèves utilisent bien, pour la plupart des variables motivationnelles, un groupe de référence auquel ils se comparent explicitement ou implicitement pour évaluer leur degré de motivation, concept de soi ou anxiété. Le profil de résultats contrastés entre la Norvège et la Belgique avant et après standardisation est quant à lui une indication forte qu'un même phénomène psychologique (se comparer à un groupe pour s'autoévaluer) produit des effets différents selon le macrocontexte (structure du système éducatif). Si le BFLPE est universel, il se décline différemment selon le degré d'homogénéisation des groupes classe et école, variable selon les systèmes éducatifs.

Van de Gaer, Grisay, Schulz & Gebhardt (2012) ont également exploré l'existence d'un effet du groupe de référence, pour rendre compte de la relation paradoxale entre attitudes et performances déjà évoquée précédemment (voir figure 1, p. 30). Leur étude porte sur les liens entre le concept de soi et les performances en sciences (PISA 2006), investiguée *via* des analyses multiniveaux (trois niveaux : élèves, écoles, pays).

À l'intérieur des écoles, les élèves les plus performants ont, sans surprise, un meilleur concept de soi en sciences. Au niveau écoles, un effet négatif de la moyenne des performances de l'école sur le concept de soi est observé ; ceci correspond typiquement à un « *big-fish-little pond effect* » (BFLPE) ; à compétences en sciences égales dans le test PISA, les élèves qui fréquentent une **école moins performante** ont un **meilleur concept de soi** en sciences. Ceci peut paraître paradoxal à première vue, mais s'explique aisément sur le plan psychologique : un élève moyen en sciences qui fréquente une école très forte se sentira comparativement moins fort que s'il fréquente une école moyenne, voire peu performante.

Cet effet négatif de la moyenne des performances de l'école sur le concept de soi est d'autant plus marqué que le pays compte une proportion élevée d'écoles sélectives (forte variance ou ségrégation entre écoles), autrement dit quand le système est fortement différencié. Ce résultat rejoint ce que Lafontaine et Monseur (2007) ont également montré : l'effet du groupe de référence varie selon le degré de différenciation des groupes classe et école du système éducatif.

Le *Big-fish-little-pond-effect* joue donc d'autant plus les catalyseurs de différences que le système est lui-même peu inclusif.

6. CONCLUSIONS

Au terme de cet examen de l'évolution de la manière dont les contextes ou leurs effets sont pris en compte dans les études comparatives à large échelle, un certain nombre de lignes de force conclusives se dégagent en termes d'acquis ou d'avancées, de soucis ou points d'attention et de défis pour l'avenir non seulement de ces enquêtes, mais aussi pour les méthodologies de l'évaluation en éducation.

6.1. Les acquis

Il ne fait aucun doute que la préoccupation pour les contextes a été croissant au cours des 20 dernières années. Les cadres conceptuels présidant aux enquêtes internationales se sont précisés au fil des cycles et sont marqués par un notable souci d'ancrer les variables à collecter dans des modèles ou des concepts validés par la recherche. Des efforts considérables ont été consentis pour développer des questionnaires de contexte de qualité. Ceci ne suffit toutefois pas encore à garantir l'équivalence entre pays pour une série de constructs/variables non-cognitifs et contextuels. Ce défi subsiste en effet au-delà de la rigueur conceptuelle et de l'application des principes qui garantissent les qualités psychométriques de base (validité et fidélité mesurées *via* les alphas et les analyses factorielles).

6.2. Les soucis ou points d'attention

Les biais de « méthodes » liés notamment aux échelles de Likert, abondamment utilisées dans les questionnaires de contexte, étaient connus de longue date : acquiescence, réponses extrêmes ou intermédiaires, désirabilité sociale, comportements satisuffisants, biais de modestie… À partir des années 2005, différentes analyses secondaires menées sur les données PISA font apparaître que ces biais ou styles de réponse sont davantage marqués dans certains pays ou cultures, ce qui explique largement le paradoxe attitudes-performances fréquemment observé. Une préoccupation de plus en plus forte se dessine alors pour les questions de stabilité cross-culturelle qui constituent une menace pour la validité des comparaisons entre pays sur les échelles concernées. Cette préoccupation est désormais bien présente au sein des organismes et principaux centres de recherche qui sont à l'initiative ou mettent en œuvre les grandes enquêtes internationales. Toutefois, les précautions qui découlent de cette prise de conscience sont peu explicitées – c'est le plus souvent dans les rapports techniques qu'elles figurent – et du chemin reste donc à parcourir pour que tous les utilisateurs des résultats des enquêtes internationales soient avertis des limites que la présence de cette relative instabilité cross-culturelle de certaines échelles des questionnaires de contexte impose aux conclusions que l'on peut en tirer.

Si l'on examine les rapports internationaux les plus récents édités par l'OCDE et l'IEA, on peut constater que dans PISA, à partir de 2012, l'OCDE applique la politique suivante : les moyennes par pays pour les variables de contexte ne sont présentées que si la corrélation avec le rendement va dans le même sens au niveau individuel ou au niveau du pays, autrement dit quand il n'y a pas de paradoxe attitudes-performances. Dans les rapports de l'IEA, la politique reste variable selon les études. Alors que le rapport de l'étude ICILS 2013 (Fraillon, Ainley, Schulz, Friedman, & Gebhardt, 2014) se montre prudent, dans le rapport de l'enquête PIRLS 2011

(Mullis, Martin, Foy, & Drucker, 2012), non seulement les moyennes des scores par pays sur certaines variables contextuelles sont comparées, mais elles le sont sous forme de « ranking ». On trouve ainsi un tableau présentant les pays classés par ordre descendant du pourcentage d'élèves qui disent ne pas aimer la lecture. Y figurent en tête, avec plus de 20 % des élèves déclarant ne pas aimer la lecture, l'Irlande du Nord, les États-Unis, la Hongrie, la Finlande, la Pologne, la Suède, le Danemark, les Pays-Bas… Dans ce groupe, on aura reconnu au passage des pays très performants en lecture. En bas du classement, avec moins de 10 % d'élèves déclarant ne pas aimer la lecture, figurent le Portugal, la Géorgie, l'Iran, l'Indonésie, le Botswana, le Maroc, l'Azerbaïdjan, l'Arabie saoudite, Oman… Le pattern maintes fois décrit dans le présent article se retrouve quasi à l'identique : ce sont les pays les moins performants, les moins riches, pays du sud ou en voie de développement qui déclarent être les plus motivés par la lecture. La présentation de ces données sans réserve, sous forme de classement, n'est évidemment pas de nature à mettre en garde contre les conclusions qu'il faut se garder de tirer ; au contraire elle y incite. Quoi de plus étonnant si certains pays, tout en s'inquiétant de leurs performances, se félicitent – et cela s'est vu – du niveau de motivation et d'engagement de leurs élèves dans la lecture… alors qu'il s'agit essentiellement d'un biais de méthode ?

6.3. Les défis

Un des défis consiste donc, maintenant que ces effets sont connus (biais/style de réponse, cadre de référence et BFLPE), à tirer toutes les conséquences de ces limites méthodologiques dans les rapports internationaux, mais aussi nationaux : la non-équivalence de certaines échelles peut conduire à des erreurs d'interprétation majeures telles que conclure que les élèves marocains sont davantage motivés par la lecture que les élèves finlandais.

Les études qui ont décrit les procédures *a priori* et *a posteriori* permettant de prévenir les biais ou d'ajuster les données après leur collecte sont encourageantes. On voit qu'il existe des techniques efficaces permettant de contrer ou de corriger ces biais. Des défis d'ordre conceptuel et méthodologique n'en subsistent pas moins. Que mesure-t-on réellement lorsque l'on applique certains ajustements (en particulier pour les vignettes d'ancrage) ? Jusqu'où faut-il ajuster ? Dans quelle mesure le biais ou style de réponse est-il une nuisance ou un mode de communication (Van de Vijver & He, 2014) ? Les MG-CFA sont-elles la façon la plus adéquate d'appréhender la stabilité cross-culturelle dans les enquêtes à très large échelle ? Le débat sur ces questions est loin d'être clos.

Même si, comme annoncé d'emblée, les questions de la juste mesure des effets de contexte ont été abordées dans ce chapitre au travers des

enquêtes internationales, qui représentent le meilleur observatoire pour ce faire, il serait totalement erroné de penser que seules celles-ci sont concernées par ces questions et ces défis.

Ces connaissances développées pour les évaluations à large échelle valent aussi à l'échelle nationale bien sûr, mais aussi pour des enquêtes à plus petite échelle voire pour des interviews qualitatives. Les variables « à haut risque » pour les styles de réponse et la mobilisation de cadres de référence engendrant le BFLPE sont les échelles de Likert, que l'on demande un degré d'accord, une fréquence temporelle, un volume, un nombre..., surtout si elles impliquent une auto-évaluation mettant en jeu une comparaison implicite ou explicite avec un groupe de référence.

Lorsque vous êtes confrontés, dans des études plus locales, au fait que les élèves des sections professionnelles, de l'enseignement spécialisé, en retard scolaire, rapportent sur des échelles de type Likert des attitudes PARADOXALES, comme une motivation ou un engagement relativement positifs par rapport à l'école ou à des matières scolaires, vous êtes vraisemblablement en présence de biais de réponse et/ou d'un BFLPE (cadre de référence). C'est le caractère paradoxal, voire incompréhensible du résultat qui doit d'abord alerter.

Il existe toujours une possibilité d'interaction entre un item, une question, un test, une méthode, un instrument de mesure et certaines caractéristiques des répondants – sexe, origine socioéconomique, culturelle ou ethnique, leur pays ou leur culture. Ce principe de base en méthodologie de l'évaluation étant connu, le défi est d'une part de trouver un *juste* équilibre en variant les modalités d'évaluation, en sorte que la balance ne penche pas toujours dans le même sens, d'autre part d'être vigilant et de toujours *remettre les données dans leur contexte* quand vient le moment de l'interprétation.

RÉFÉRENCES BIBLIOGRAPHIQUES

Arends-Tóth, J., Van de Vijver, F. J. R., & Poortinga, Y. H. (2006). The influence of method factors on the relation between attitudes and self-reported behaviors in the assessment of acculturation. *European Journal of Psychological Assessment, 22* (1), 4-12.

Bonnet, G. (2002). Reflections in a critical eye : on the pitfalls of international assessment. *Assessment in Education, 9*, 387-400.

Bressoux, P. (2013). *Modélisation statistique appliquée aux sciences sociales*. Bruxelles : De Boeck Supérieur.

Buckley, J. (2009). *Cross international response styles in international educational assessments : Evidence from PISA 2006*. Last accessed [10/10/2012] http://polmeth.wustl.edu/workingpapers.php

Chen, C., Lee, S., & Stevenson, H.W. (1995). Response style and cross-cultural comparisons of rating scales among East Asian and north American students. *Psychological Sciences, 6* (3), 170-175.

Fraillon, J., Ainley, J., Schulz, W., Friedman, & Gebhardt, E. (2014). *Preparing for life in a digital age. The IEA international Computer and Information Literacy Study*. Springer.

Goldstein H. (November 2004). International comparisons of student attainment : some issues arising from the PISA study. *Assessment in Education, Vol. 11, 3,* 319-330.

Grisay, A. (2009). *Draft note on social desirability in PISA 2006*. Note inédite

Grisay, A. (2010). *Missing date issues and response styles problems in PISA 2009 student questionnaire. Draft note to the TAG*. Melbourne 15-18 mars. Note inédite.

Harzing, A.W. (2006). Response styles in cross-national survey research : A 26-country study. *International Journal of Cross cultural Management, 6* (2), 243-266.

Heine, S.J., Kitayama, S., Lehman, D.R., Takata, T., Ide, E., Leung, C., & Matsumoto, H. (2001). Divergent consequences of success and failure in Japan and North America : An investigation of self-improving motivations and malleable selves. *Journal of Personality and Social Psychology, 81* (4), 599-615.

Hofstede, G. (2001). *Culture's consequences.* Beverly Hills : Sage.

Krosnick, J. (1991). Response strategies for coping with cognitive demands of attitude measures in surveys. *Applied cognitive psychology, 5,* 213-236.

Kyllonen, P. & Bertling, J. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier & D. Rutkowski (eds). *Handbook of international large-scale assessment* (pp. 277-285). Londres : Chapman & Hall.

Lafontaine, D., & Demeuse, M. (2002). Le bon (critique), la brute (médiatique) et les truands (anglo-saxons). (Controverse. L'enquête OCDE sur les acquis des élèves en débat). *Revue Nouvelle, 115* (3-4), 100-108.

Lafontaine, D. & Monseur, C. (2007). Why do non-cognitive variables better predict mathematics achievement in some countries than in others ? A methodological study on PISA 2003. *Communication à la 12th Biennal Conference of the European Association for Research on Learning and Instruction (EARLI)*, Budapest, août 2007. http://hdl.handle.net/2268/120412.

Lafontaine, D., Baye, A., Vieluf, S., & Monseur, M. (2015). Equity in opportunity-to-learn and achievement in reading : A secondary analysis of PISA 2009 data. *Studies in Educational Evaluation, 47,* 1-11.

Lie, S., & Turmo, A. (2005). *Cross-country comparability of students'self-reports-Evidence from PISA 2003*. OECD/PISA, Document : TAG (0505) 11.

Marsh, H., & Hau, K. (2003). Big-fish-little-pond effect on academic self-concept : A cross-cultural (26 countries) test of the negative effect of academically selective schools. *American Psychologist, 58* (5), 364-376.

Marsh, H., Tau, K.-T., Artelt, C., Baumert, J. & Peschar, J. (2006). Oecd's brief self-report measure of educational psychology's most useful affective constructs : cross-cultural, psychometric comparisons across 25 countries. *International Journal of Testing, 6* (4), 311-360.

Marsh, H., & O'Hara, (2010). Long-term total negative effect of school-average ability on diverse educational outcomes : direct and indirect effects of the big-fish-little-pond effect. *Zeitschrift für Pädagogische Psychologie. 24,* 51-72.

Marsh, H., Kuyper, H., Morin, A. J. S., Parker, P., & Seaton, M. (2014). Big-fish-little-pond social comparison and local dominance effects : Integrating new statistical models, methodology, design, theory and substantive implications. *Learning and Instruction, 33,* 50-66.

Mullis, I.V.S., Martin, M.O., Foy, P., & Drucker K.T. (2012). *PIRLS 2011 International Results in Reading.* TIMSS & PIRLS International Study Center, Boston College : Chestnut Hill, MA, USA & IEA, Amsterdam.

Nagengast, B. & Marsh, H. (2012). Big fish in little ponds aspire more : mediation and cross-cultural generalizability of school-average ability effects on self-concept and career aspirations in science. *Journal of Educational Psychology,* Advance online publication. doi : 10.1037/a0027697

OCDE (1999). *Mesurer les connaissances et compétences des élèves. Un nouveau cadre d'évaluation*. Paris : OCDE.

OCDE (2003). *Cadre d'évaluation de PISA 2003 – Connaissances et compétences en mathématiques, lecture, sciences, résolution de problèmes*. Paris : OCDE.

OCDE (2006). *Compétences en sciences, lecture et mathématiques. Le cadre de référence de PISA 2006*. Paris : OCDE.

OCDE (2009). *Le cadre d'évaluation de PISA 2009. Les compétences clés en compréhension de l'écrit, en mathématiques et en sciences*. Paris : OCDE.

OCDE (2012). *Cadre d'évaluation et d'analyse du cycle PISA 2012. Compétences en mathématiques, en compréhension de l'écrit, en sciences, en résolution de problèmes et en matières financières*. Paris : OCDE.

OCDE (2014). *PISA 2012 Technical report*. Paris : OCDE.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research : a critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88* (5), 879-903.

Romainville, M. (March-April 2002). Du bon usage de PISA. *La Revue Nouvelle, 3-4*, 86-99.

Schulz, W. (2005). *Testing parameter équivalencefor questionnaire indices using confirmatory factor analysis and item response theory.* Paper presented at the Annual Meeting of the American Educational Research Association in San Francisco, 7-11 April, 2005.

Seaton, M. Marsh, H. & Craven, R. (2010). Earning its place as a pan-human theory : universality of the big-fish-little-pond-effect across 41 culturally diverse countries. *Journal of educational psychology, 101* (2), 403-419.

Trautwein, U., Lüdtke, O., Marsh, H. & Nagy, G. (2009). Within-school social comparison : how students perceive the standing of their class predict academic self-concept. *Journal of educational psychology, 101* (4), 853-866.

Van de Gaer, E., Grisay, A., Schulz, W., & Gebhardt, E. (2012). The Reference Group Effect : An Explanation of the Paradoxical Relationship between Academic Achievement and Self-Confidence across Countries. *Journal of Cross-cultural Psychology,* http://jcc.sagepub.com/content/early/2012/02/08/0022022111428083

Van Herk, H., Poortinga, Y.H., & Verhallen, T.M.M. (2004). Response styles in rating scales : Evidence of method bias in data from 6 EU countries. *Journal of Cross-Cultural Psychology, 35* (3), 346-360.

Van de Vijver, F. J. R., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment : An overview. *Revue européenne de psychologie appliquée/European Review of Applied Psychology, 54*,119-135.

Van de Vijver, F. J. R., & He J. (2014). *Report on social desirability, midpoint and extreme responding in TALIS 2013*. OECD Education Working Papers, n° 107. Paris, France : OECD Publishing.

Yang, Y., Harkness, J. A., Chin, T.-Y, & Villar, A. (2010). Response Styles and Cultures. In J.A. Harkness, M. Braun, B. Edwards, T.P. Johnson, L. Lyberg, P. Ph. Mohler, P.-E. Pennell, & T. W. Smith (Eds.). *Survey Methods in Multinational, Multiregional and Multicultural Contexts* (pp. 203-233). Hoboken : Wiley & Sons.