

Conférence inaugurale

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

Dominique Lafontaine

Département Éducation et Formation - Université de Liège

Évaluations à large échelle

Les évaluations comparatives à large échelle sont un objet privilégié pour étudier la prise en compte et la mesure des effets de contexte pertinents ou indésirables (biais).

Des contextes multiples

Dans les évaluations comparatives à large échelle, les élèves dont on mesure les performances scolaires ou cognitives

- ▶ sont issus de langues, de cultures, d'horizons et de systèmes éducatifs divers ;
- ▶ sont nichés dans des classes, des filières, des programmes d'études, des écoles, des régions, des pays (= contextes emboîtés).

Comment cette diversité de contextes est-elle prise en compte ?

Point de départ/posture

Les études comparatives à large échelle (ELA):

- ▶ ne sont pas la mesure de toute chose... ;
- ▶ ont toute leur pertinence, **à condition de tenir compte des contextes, en amont et en aval de l'évaluation.**

Comparer des pommes et des poires



Comparer fait sens...

- ▶ Pour autant qu'on s'intéresse à l'arbre, au verger, au sol, aux conditions climatiques, aux traitements phyto-sanitaires... qui permettent de produire des fruits divers, non formatés, dans le respect de l'environnement ...

Angle d'approche

- ▶ En posant d'une part que le cadre des ELA est pertinent, d'autre part que l'importance des contextes est capitale,
- ▶ il s'agit de revisiter un vaste corpus d'ELA
- ▶ d'y situer 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 ELA.

Principales études à large échelle

	1960	1970	1980	1990	2000	2010
Langue maternelle		Reading, Literature	Composition	RLS IALS (1998)	PIRLS 2001, 2006 Pisa2000, 2003, 2006, 2009 ALL 2003, 2006	PIRLS 2011, 2016 Pisa 2012, 15, 18 PIACC 2010
Math	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
Science		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

Prendre la juste mesure des effets de contexte

1. En amont du test/questionnaire

- Cadres de référence conceptuels (*Frameworks*)

2. Dans le questionnaire (validation)

- Invariance/stabilité cross-culturelle
- Biais de méthodes/styles de réponses

Prendre la juste mesure des effets de contexte

3. En aval du questionnaire (résultats et interprétation)

- Influence des groupes de référence et du macro-contexte “système éducatif”.

Prendre la juste mesure des effets de contexte rend la comparaison éclairante et évite de comparer l'incomparable (par défaut de validité).

En amont des questionnaires

Cadres de référence conceptuels (*Frameworks*)



Cadres de référence : penser les effets de contexte

Oecd (1999), *PISA 2000 Measuring students' knowledge and skills*.
Paris: Ocdé.

- ▶ Le principal focus est mis sur la mesure cognitive (savoirs et compétences) : une seule page (sur 85) est consacrée aux questionnaires de contexte
- ▶ L'importance "d'indicateurs contextuels mettant en relation les performances avec les caractéristiques des élèves et des écoles" (Oecd, 1999, p. 10) est soulignée...

Cadres de référence PISA

- ▶ mais pas de cadre de référence distinct.
- ▶ Aucune trace reconnue d'un groupe d'experts réfléchissant aux questionnaires de contexte.
- ▶ Pas d'évolution majeure en 2003, 2006; le QEG (T. Williams, ACER) travaille dans l'ombre. "The questionnaire framework was not published by the OECD but is available as a project working document" (*Technical report PISA 2000*, p. 34).

POINT de RUPTURE = 2009

- ▶ En 2009, un chapitre intitulé Questionnaire framework apparaît (20 pages).
- ▶ Conceptualisation systématique des constructs essentiels au niveau système, école, classe, élève.
- ▶ En annexe, liste systématique des constructs et des questions/items les mesurant.

Vers un cadre distinct : PISA 2015

- ▶ Élaboration d'un cadre de référence propre aux Questionnaires de contexte: Oecd (2013).

The PISA 2015 Framework for Context Assessment. Monitoring Opportunities and Outcomes, Policies and Practices. Modelling Patterns and Relations, Impacts and Trends in Education. Draft (Field Trial) version, October 2013. Produced by Core 6 (Dipf, Frankfurt-am-Main).

En résumé : de PISA 2000 à 2015

- ▶ Montée en puissance des questionnaires de contexte
- ▶ Reconnaissance du nécessaire travail sur les fondements conceptuels (framework et QEG) (2009)
- ▶ Souci croissant pour la qualité, la validité et la comparabilité cross-culturelle des variables contextuelles ou non cognitives (2012 et 2015)

Dans le questionnaire

- Invariance/stabilité cross-culturelle
- Biais de méthode/styles de réponses

Effets de contexte “indésirables”/biais

- ▶ Étapes de validation lors de l’essai de terrain
→ repérer/détecter les effets de contexte et les biais/styles de réponse qui peuvent menacer la validité, la stabilité ou l’équivalence des données.

Biais de méthode

POINT de rupture : 2009/2012

- ▶ Un manque de comparabilité peut résulter de la « méthode » utilisée, par ex. l'échantillonnage, le mode d'administration, ou l'instrument lui-même (Podsakoff, MacKenzie, Lee, & Podsakoff, 2003; van de Vijver & Tanzer, 2004).
- ▶ Prise de conscience (recherches en psychologie cross-culturelle) des biais de réponse systématiques engendrés en particulier par les mesures auto-rapportées (échelles de Likert).
- ▶ Les styles de réponse ont une composante culturelle.

Biais de méthode

- ▶ Plusieurs analyses secondaires menées sur des ELA et PISA mettent en évidence des biais de « méthode » pour les mesures auto-rapportées (échelles de Likert): Buckley (2009); Grisay (2009; 2010); Lie & Turmo (2005), Schlutz (2005); Van de Gaer, Grisay, Schulz, Gebhardt (2012).

Biais de méthode/styles de réponse

- ▶ Sur des échelles telles que celles mesurant l'intérêt pour la lecture ou les maths, des biais de réponse (désirabilité sociale, tendance à choisir les extrêmes ou les intermédiaires, comportements "satisfaisants") sont observés; ils sont plus marqués dans certains pays ou zones géographiques.
- ▶ Biais de modestie dans les cultures asiatiques, davantage de choix des extrêmes dans les cultures occidentales; acquiescence, désirabilité sociale plus marquées dans les pays du sud ((Harzing, 2006; van Herk, Poortinga, & Verhallen, 2004; Yang, Harkness, Chin, & Villar, 2010)
- ▶ Ceci constitue une menace pour la validité : la mesure du construct ciblé (par ex: le concept de soi) est "polluée" par une tendance à réagir à l'instrument (échelle de Likert) variable selon les sujets répondants, mais aussi les cultures ou les pays.

Biais de méthode

- ▶ Il en résulte une (certaine) absence d'équivalence et de comparabilité des réponses
- ▶ Ceci conduit notamment au **paradoxe Attitudes-Performances** : la corrélation entre les attitudes/performances est positive dans les pays, mais négative quand on calcule les corrélations au niveau pays, car les pays les moins performants sont ceux qui déclarent les attitudes les plus positives.
- ▶ Ce phénomène a été observé à de nombreuses reprises dans les ELA (Kennedy & Trong, 2006; Kyllonen & Bertling, 2013; Lie & Turmo, 2005; Shen & Tam, 2008; van de Gaer, Grisay, Schulz, & Gebhardt, 2012; Wilkins, 2004).

Comment repérer les biais de méthode ?

- ▶ L'absence d'invariance (stabilité cross-culturelle) des échelles se teste en appliquant des MG-CFA (Multiple Group Confirmatory Factorial Analyses) ou des MG-SEM (Nagengast, & Marsh, 2014).
- ▶ Trois types d'invariance sont testées :
 - Invariance configurale : même structure factorielle dans les différents pays,
 - Invariance métrique : nécessaire pour que les facteurs puissent être corrélés avec d'autres variables de façon valide, comparable d'un pays à l'autre;
 - Invariance scalaire : nécessaire pour pouvoir comparer les moyennes des facteurs d'un pays à l'autre (= la plus difficile à atteindre).

Comment prendre en compte les biais de méthode?

(neutraliser les effets de contexte)

- ▶ **Dès la conception du questionnaire** : utiliser des formats d'items alternatifs.
- ▶ Dans PISA 2012 et 2015 (FT), une série d'alternatives ont été explorées, en exploitant les possibilités offertes par l'administration par ordinateur.

Format d'items alternatifs : choix forcés

- ▶ Pour certaines échelles portant sur les stratégies, à côté du format habituel (Likert scale), 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 :
de -0.47 (Likert) à + 0.60 (choix forcé) (Kyllonen, & Berkling, 2013).

Items « piège » (signal detection correction for topic familiarity)

- ▶ OTL : une liste de 19 termes mathématiques est fournie aux élèves (logarithme, probabilités, cosinus, aire...) parmi lesquels on a glissé quelques termes fictifs (échelle subjonctive, nombre propre, fraction déclarative...); degré de familiarité (de « jamais entendu » à « je connais et je comprends le concept »).

- ▶ Deux scores sont calculés :
 - 1) Score avec tous les concepts OTL
 - 2) Score ajusté en soustrayant la moyenne pour les faux concepts de la moyenne pour tous les concepts.

Items « piège » (signal detection correction for topic familiarity)

- ▶ Corrélations des deux scores de familiarité avec le score en maths dans chaque pays = 0.45 et 0.44.
 - ▶ Corrélations au niveau pays : score total = 0.17, score ajusté = 0.54
- **Dans certains pays, les élèves ont tendance à indiquer plus souvent qu'ils connaissent des concepts mathématiques inexistants (acquiescence, désirabilité sociale ou comportement satisfaisant).**

Vignettes d'ancrage

- ▶ Cette technique consiste à faire évaluer une série de vignettes dites communes et à s'en servir pour ajuster les scores d'autres échelles.

Vignettes d'ancrage

PISA 2012 : trois vignettes d'ancrage ont été créées en lien avec l'échelle relative au soutien des enseignants

1. *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.*
2. *M. Guesse donne des devoirs de maths une fois par semaine. Il rend toujours les devoirs corrigés avant les examens. M. Guesse se soucie des apprentissages de ses élèves.*
3. *Mme Stevens donne des devoirs de maths une fois par semaine. Elle ne rend jamais les devoirs corrigés avant les examens. Mme Stevens se soucie des apprentissages de ses élèves.*

Vignettes d'ancrage

- ▶ Les vignettes d'ancrage servent à détecter la tendance du répondant à juger de manière modérée ou extrême les vignettes communes (profs hypothétiques) et comment il discrimine les items.
- ▶ Les trois vignettes d'ancrage servent pour ajuster l'échelle relative au soutien de l'enseignant (*Mon prof de maths fournit de l'aide supplémentaire si nécessaire... : 4 items avec 4 degrés d'accord*).

Vignettes d'ancrage : Kyllonen, & Berkling, 2013

- ▶ 63 pays participants

- ▶ Corrélations entre l'échelle de soutien par l'enseignant et le score en maths
 - 1. dans chaque pays :**
 - Score non ajusté = **0.03**
 - Score ajusté grâce aux vignettes = **0.13**
 - 2. au niveau pays :**
 - score non ajusté = - **0.45**
 - score ajusté = + **0.29**

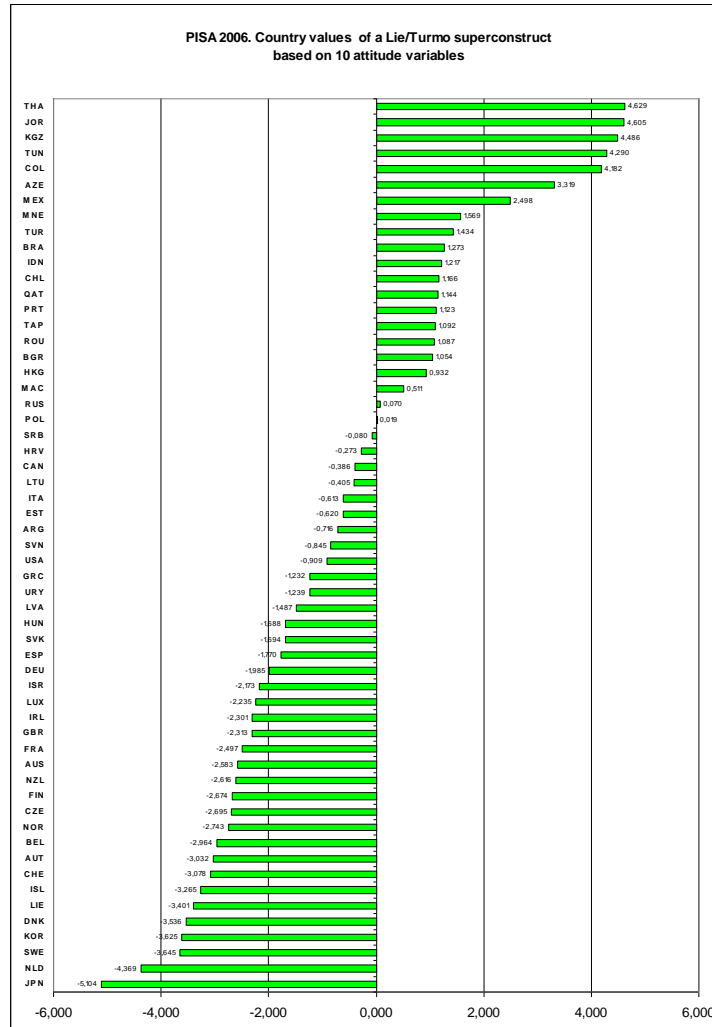
Comment prendre en compte les biais de méthode?

(neutraliser les effets de contexte)

- ▶ **Après coup:** modélisations statistiques pour tenter d'isoler et ensuite corriger la tendance des sujets ou des pays à donner des réponses socialement désirables (Podsakoff, & al., 2003).
- ▶ Lie, & Turmo (2005) ont appliqué une analyse factorielle à 10 échelles de Likert dans PISA 2003; facteur principal sur lequel tous les items ont des saturations élevées ((« superconstruct »)).
- ▶ Des scores élevés du pays sur le facteur principal indiquent que les élèves de ce pays ont tendance à donner des réponses fortement positives quel que soit le construct mesuré (self-concept, relations élèves-enseignants, coopération et compétition, attitudes envers l'école...)

Lie & Turmo (2005):

Moyennes des pays sur le superconstruct (10 échelles de Likert)



Comment prendre en compte les biais de méthode?

(neutraliser les effets de contexte)

- ▶ Très haut sur le superconstruct : Thaïlande, Jordanie, Tunisie, Colombie, Azerbaïdjan, Mexico, Montenegro, Turquie, Brésil, Indonésie, Chili, Qatar, Portugal, Taipei, Roumanie...
- ▶ Tout en bas sur le superconstruct : Japon, Pays Bas, Suède, Corée, Danemark, Liechtenstein, Suisse, Autriche, Belgique, Norvège, Tchéquie, Finlande, Nouvelle Zélande, Australie, France...
- ▶ Le superconstruct explique 66% de la variance.
- ▶ En utilisant le superconstruct pour ajuster le score des pays sur les échelles de Likert, les corrélations deviennent nettement moins négatives, voire positives : par ex. pour le concept de soi en maths de -0.26 à $+0.74$!

Lafontaine, Baye, Vieluf, & Monseur (2014)

- ▶ **Opportunities-to-learn and reading achievement in a cross-cultural context**
- ▶ Variables OTL lecture PISA 2009 (17 items, échelle de Likert)
- ▶ CFA : 4 facteurs cohérents, faiblement, certains négativement corrélés avec le rendement.
- ▶ Un biais de méthode/style de réponse est suspecté.

Lafontaine, Baye, Vieluf, & Monseur (2014)

- ▶ Analyse IRT (within-item) tridimensionnelle est appliquée
- ▶ Dimension 1= tous les items OTL (Hypothèse : capte ce qui est commun à tous les items, dont le biais de méthode ou style de réponse).
- ▶ Deux dimensions spécifiques OTL mesurant le construct “épuré” du biais de réponse.

Corrélations entre les trois dimensions et le score en lecture

	Common method	OTL textes non continus	OTL fiction
Moyenne Ocdé	- 0.38	0.43	0.52

Lafontaine, Baye, Vieluf, & Monseur (2014)

- ▶ Corrélations fortes de la dimension Common factor avec deux indices *acquiescence* et *comportement satisfaisant* → biais/style de réponse.
- ▶ Les deux indices de biais sont eux-mêmes très négativement reliés avec le score en lecture → les pays peu performants présentent davantage de styles de réponse biaisés.

Biais de méthode/styles de réponse

- ▶ Des biais/styles de réponse affectent bien certaines des échelles d'une manière systématique et non aléatoire (patterns "culturels" et autres – genre, niveau d'aptitude)
- ▶ 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).

En aval du questionnaire

Analyses, résultats, interprétation et politiques éducatives :

la juste place des contextes

Comprendre et interpréter

- ▶ Tenir compte des micro-contextes (classe, école) qui constituent des groupes de référence (*frame/group of reference*), cf. théories de la comparaison sociale;
- ▶ Tenir compte du macro-contexte système éducatif.

Études empiriques

Lafontaine & Monseur (2007), *Why do non-cognitive variables better predict mathematics achievement in some countries than in others? A methodological study of PISA 2003.* Communication à Earli Budapest.

Lafontaine & Monseur (2007)

- ▶ Variables non cognitives liées aux mathématiques (self-concept, motivation extrinsèque et intrinsèque, auto-efficacité, anxiété).
- ▶ Corrélations avec le rendement : 0.36 en moyenne, mais grande variabilité selon les pays : pour le self-concept, 0.22 en Belgique vs 0.56 en Norvège.
- ▶ Variation non aléatoire (patterns) : corrélations plus élevées dans les systèmes éducatifs compréhensifs → variation liée au contexte.

Lafontaine & Monseur (2007)

Hypothèses

▶ Double effet de contexte

1. 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);
2. 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 (BFLPE) se traduit différemment selon le contexte 2 (système éducatif).

Lafontaine & Monseur (2007)

Hypothèses

- ▶ Dans les systèmes éducatifs compréhensifs (pas de filières, pas de redoublements, variance entre écoles faible), les groupes de référence servant de base de comparaison diffèrent peu entre eux.
- ▶ Dans les systèmes différenciés, par définition, les groupes de référence ont des profils académiques très contrastés (écoles/classes très ou très peu performantes). Le point de comparaison (norme) que prennent les élèves pour s'auto-évaluer y est donc éminemment variable, et les effets liés au groupe de référence sont démultipliés.
- ▶ L'effet groupe de référence ne sera pas observé pour self-efficacy (pas de groupe de référence).

Lafontaine & Monseur (2007)

Méthodes

- ▶ Pour tester les hypothèses, les variables motivationnelles ont été standardisées par école/filière/classe d'étude (variables centrées et réduites);
- ▶ Les corrélations avant/après standardisation ont été comparées; la standardisation neutralise le côté variable du groupe de référence (écarts à la moyenne du groupe plutôt que valeurs absolues).

Corrélations avant/après standardisation par école/filière/année d'étude

		<i>Motivation intrinsèque</i>	<i>Motivation extrinsèque</i>	<i>Concept de soi</i>	<i>Anxiété</i>	<i>Auto efficacité</i>
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 Ocdé	Brute	0.14	0.12	0.35	-0.34	0.51
	Standard	0.21	0.17	0.37	-0.33	0.40

Lafontaine & Monseur (2007)

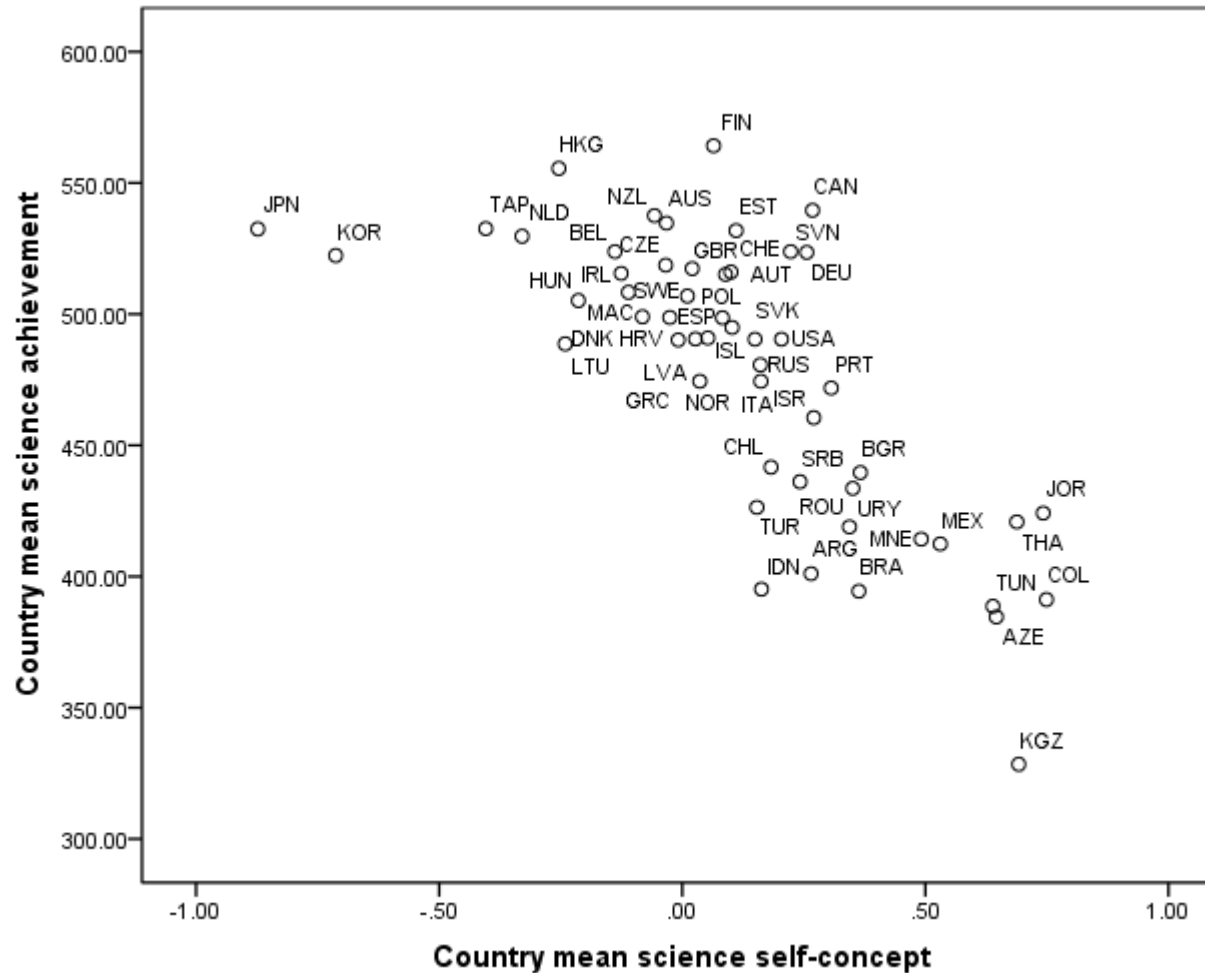
Résultats

- ▶ Les élèves utilisent bien pour certaines variables motivationnelles un groupe de référence (BFLPE).
- ▶ Ce qui produit des effets différents selon le contexte (structure du système éducatif).

Van de Gaer, E., Grisay, A., Schlutz, W., Gebhardt, E. (2012)

- ▶ Ces auteurs ont également exploré l'existence d'un effet du groupe de référence, pour rendre compte de la relation paradoxale entre attitudes et performances.
- ▶ Analyses multiniveaux (élèves, écoles, pays), données PISA 2006.

Van de Gaer, & al., 2012



Van de Gaer, & al., 2012

- ▶ À l'intérieur des écoles, les garçons, les non-natifs, les élèves plus privilégiés ont un meilleur self-concept en sciences (à compétences égales); les élèves en retard scolaire ont un meilleur self-concept
- ▶ Au niveau écoles, un effet négatif de la moyenne des performances de l'école sur le self-concept est observé (BFLPE); dans les écoles moins performantes, les élèves ont un meilleur concept de soi en sciences (PARADOXE).
- ▶ Cet effet négatif de la moyenne de l'école est d'autant plus marqué que le pays compte une proportion élevée d'écoles sélectives (forte variance ou ségrégation entre écoles).

Le BFLPE :

se sentir un gros poisson dans la petite mare (effet de contexte)

- Théorie de la comparaison sociale : l'auto-évaluation dépend du contexte et des groupes dont on fait partie. Travaux de Herb Marsh et collaborateurs.
- Concept de soi académique (ASC « academic self-concept »): *je suis fort en maths, en sciences...* (par rapport à, au moins implicite).

Le BFLPE : un effet généralisé, universel selon Marsh

- Les effets sont robustes et observés dans une série de contextes, en utilisant notamment les données de PISA (Marsh & Tau, 2003; Seaton, Marsh, Craven, 2010);
- Ils concernent parfois d'autres variables que le concept de soi comme les aspirations d'études (Nagengast & Marsh, 2012), l'intérêt, le choix d'études (Marsh & O'Hara, 2010).

Le BFLPE : un effet de contexte généralisé, pour toutes les mesures auto-rapportées

▶ 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'auto-évaluer ou rendre compte de ses pratiques, et que pour ce faire, il utilise potentiellement un groupe de référence.

▶ Auto-évaluation de la taille/poids: comment chacun juge-t-il sa taille et son poids ?

-Très petit/plutôt petit/plutôt grand/très grand

- Selon qu'il est un homme/une femme, Hollandais ou Vietnamien, issu d'une famille de grands ou de petits?

Conclusions

Acquis, soucis...

....défis

Conclusions : les acquis

- ▶ La préoccupation pour les contextes a été croissant au cours des 20 dernières années.
- ▶ Cadres conceptuels de mieux en mieux définis, souci d'ancrer les variables sur des modèles ou des concepts validés par la recherche.
- ▶ Efforts considérables consentis pour développer des questionnaires de contexte de qualité.
- ▶ Ce qui ne suffit pas à garantir l'équivalence entre pays pour une série de constructs/variables non-cognitifs et contextuels.

Conclusions : les soucis

- ▶ Préoccupation de plus en plus forte pour les questions de **stabilité cross-culturelle** (menace pour la validité des comparaisons entre pays).
- ▶ Biais de « **méthodes** » liés notamment aux échelles de Likert : les biais connus (acquiescence, réponses extrêmes ou intermédiaires, comportements satisfaisants, modestie...) sont davantage marqués dans certains groupes (genre, niveau d'habileté...) ou cultures.
- ▶ Influence des **groupes de référence** (BFLPE) qui conduisent au paradoxe attitudes-rendement, davantage marquée dans certains **systèmes éducatifs**.

Conclusions : les soucis

- ▶ L'importance d'établir l'invariance dans les recherches cross-culturelles est reconnue sur le plan méthodologique (Chen, 2008)
- ▶ Mais commence seulement à « percoler » dans les recherches appliquées.

Rapports internationaux récents

- ▶ Au niveau de PISA, l'Ocdé applique maintenant 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.

Rapports internationaux récents

- ▶ Pour l'IEA, la politique peut varier selon les études : dans le dernier rapport Pirls 2011, non seulement les moyennes par pays des variables contextuelles sont comparées, mais elles le sont sous forme de “ranking”.

Par ex: pays classés par ordre descendant du % d'élèves qui disent ne pas aimer la lecture

>20%: Irlande du Nord, USA, Hongrie, Finlande, Pologne, Suède, Danemark, Pays-Bas...

< 10% Portugal, Georgie, Iran, Indonésie, Botswana, Maroc, Azerbaïdjan, Arabie saoudite, Oman...

Conclusions : les défis

- ▶ Affiner les techniques/approches permettant de corriger ces biais
- ▶ Comprendre (validité) ce que l'on mesure quand on applique certains ajustements : en particulier pour les vignettes (jusqu'où ajuster ?)
- ▶ Tirer toutes les conséquences de ces limites méthodologiques dans les rapports internationaux, mais aussi nationaux : la non-invariance de certaines échelles peut conduire à des erreurs d'interprétation majeures (par ex. conclure que les élèves marocains sont davantage motivés par la lecture que les élèves finlandais).

Conclusions : à large portée

- ▶ Pas mal de variables sont concernées : l'impact n'est pas marginal
- ▶ Les variables « à haut risque » = échelles de Likert (d'accord, de temps, de volume, de nombres...), **surtout si elles impliquent une comparaison implicite ou explicite avec un groupe de référence.**

Conclusions : à transférer

- ▶ Ces connaissances développées pour les évaluations à large échelle valent aussi à plus petite échelle...
- ▶ 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, 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).
- ▶ Les mesures auto-rapportées doivent toujours être prises avec prudence; des solutions existent pour améliorer la mesure.

Conclusions : vigilance

- ▶ Il existe toujours une possibilité d'interaction entre l'item, la question, le test, la méthode, l'instrument de mesure et certaines caractéristiques des répondants – sexe, origine socioéconomique, culturelle ou ethnique, pays (possibilité d'effets croisés).

Conclusions : principe de précaution

- ▶ Utiliser de préférence d'autres formats dès la conception
- ▶ Sélectionner les concepts/échelles/items les plus stables: par exemple *self-efficacy* > *self-concept*.
- ▶ Préférer les concepts **liés aux disciplines** aux concepts généraux.
- ▶ Principe de précaution radical : ne pas comparer/présenter les moyennes quand l'invariance scalaire n'est pas atteinte ?
- ▶ Principe de précaution soft : ne pas comparer/présenter les moyennes quand un paradoxe attitudes-rendement est observé ?

Trouver la juste mesure ou l'équilibre mesuré

Défi de taille pour les comparaisons internationales, les recherches cross-culturelles, l'évaluation en général, et l'Admee en particulier !