Entre robustesse et fragilité, les résultats et classements PISA à l'aune de l'effort investi dans le test

Between robustness and fragility, the PISA results and rankings in the light of test-taking effort

Élodie Pools – elodie.pools@uliege.be
Christian Monseur – cmonseur@uliege.be – https://orcid.org/0000-0002-1357-8966
Université de Liège - Belgique

Pour citer cet article : Pools, E., & Monseur, C. (2022). Entre robustesse et fragilité, les résultats et classements PISA à l'aune de l'effort investi dans le test. *Évaluer. Journal international de recherche en éducation et formation, 8*(3), 65-92. https://doi.org/1048782/e-jiref-8-3-65

Résumé
Les enquêtes internationales en sciences de l'éducation informent les scientifiques, les décideurs politiques et l'opinion publique sur la performance de leur pays, l'exemple le plus connu étant le Programme International pour le Suivi des Acquis des élèves (PISA). Cependant, ces épreuves étant sans enjeu pour les élèves, un manque d'effort peut limiter la validité des résultats des études. Sur base des données PISA 2018 en lecture de 36 pays, cet article quantifie l'effort investi dans le test en se basant sur les temps de réponse aux items. L'impact du manque d'effort sur l'estimation d'un indicateur d'efficacité (performances moyennes) et d'équité (différences filles-garçons de performance) est analysé par le biais d'un filtrage.

Le manque d'effort dans le test entraine une sous-estimation de la performance moyenne des pays, en particulier dans ceux développant le moins d'effort. Ce manque d'effort altère les classements des pays ; cependant, les résultats globaux (en termes de performance supérieure/similaire/inférieure à la moyenne de l'OCDE) ne sont pas substantiellement impactés. Par ailleurs, les garçons étant moins engagés dans le test, les différences de moyenne et de variance en lecture entre les filles et les garçons sont surestimées.

Ces résultats soulignent l'importance de tenir compte de l'effort investi dans les épreuves à faibles enjeux. Ils pointent une des limites d'une lecture basée sur les seuls palmarès internationaux, sans toutefois remettre en cause les grandes tendances dans les résultats. Des implications pour la construction de tests et l'interprétation des résultats sont proposées.

Mots-clés
Effort dans le test ; indicateurs d'efficacité ; indicateurs d'équité ; évaluation internationale à large échelle.

Abstract

International surveys in education inform scientists, policy makers and public opinion on their country's achievement, the most famous example being the Program for International Student Assessment (PISA). However, as these tests have no stake for the students, a lack of effort might limit the validity of the studies' results. Based on the PISA 2018 reading data of 36 countries, this article quantifies test-taking effort based on item response times. The influence of a lack of effort on an efficacy (average achievement) and equity (gender differences in achievement) indicators estimate is analyzed through motivation filtering.

The lack of test-tacking effort generates an underestimation of the countries mean achievement, especially in the least effortful countries. This lack of effort distorts the country rankings; nevertheless, the overall results (in terms of mean achievement above/similar to/under the OECD average) are not substantially impacted. Furthermore, as boys are less engaged in the test, differences between boys and girls in means and variances in reading are overestimated.

These results outline the importance of accounting for test-taking effort in low-stakes assessments. They emphasize one of the limits of an interpretation based only on international league tables, while the principal trends in the results are little impacted. Implications for test construction and score interpretation are proposed.

Keywords

Test-taking effort; efficacy indicators; equity indicators; international large-scale assessment.

1. Introduction

Les enquêtes à large échelle en sciences de l'éducation produisent des indicateurs relatifs au fonctionnement des systèmes éducatifs dans une perspective internationale. Le Programme International pour le Suivi des Acquis des élèves (PISA), programme triennal mené sous l'égide de l'Organisation de Coopération et de Développement Économiques (OCDE), est sans doute la plus connue et la plus médiatisée de ces enquêtes comparatives. Dans les pays participants, il est demandé aux élèves échantillonnés de passer un test de performance portant sur différents domaines (principalement la lecture, les mathématiques et les sciences). Les réponses des élèves sont ensuite utilisées afin d'estimer la compétence de ces populations et, par la suite, de dériver différents indicateurs d'efficacité et d'équité. Ces résultats sont particulièrement attendus par la presse et sont couramment évoqués par les différents acteurs de l'enseignement, y compris lors des débats parlementaires : il est ainsi coutume, lors de la publication d'un nouveau rapport PISA, de trouver des commentaires sur ses résultats dans les journaux, les classements des pays (notamment en fonction de leur performance moyenne) y tenant régulièrement une place de choix (Cattonar & Mangez, 2014 ; Grey & Morris, 2018 ; Pons, 2010).

Si les décideurs politiques attribuent généralement des enjeux élevés à la participation de leur pays à ces enquêtes, les élèves, quant à eux, ne doivent probablement pas percevoir un intérêt majeur à y participer. En effet, il n'y aura aucune conséquence associée à leur performance au test, celle-ci ne comptant par exemple pas pour le bulletin. Une préoccupation majeure liée aux épreuves à faibles enjeux touche à la confusion dans la mesure de la performance entre l'effet de la compétence de l'élève et celui, non pertinent, de la motivation à passer le test (Eklöf, 2010 ; Finn, 2015). En effet, ces tests visent à estimer la compétence des élèves dans un domaine donné. Si l'élève ne s'engage pas dans la tâche, sa compétence risque d'être sous-estimée. La motivation à engager des efforts lors de la passation du test est donc bien un construit distinct de la compétence mesurée et constitue une source de variance non pertinente, contaminant la mesure du construit ciblé par le test, à savoir la compétence des élèves. Alors qu'une différence de performance entre deux élèves devrait ne refléter qu'une différence de compétence entre eux, elle pourrait également être le résultat d'une différence d'efforts consentis entre des élèves de compétence similaire. Plus spécifiquement, cette contamination de la mesure de la compétence traduit le « degré auquel les scores au test sont affectés par des processus étrangers à l'objectif visé par le test » (American Educational Research Association et al., 2014), l'élément contaminant étudié étant donc la motivation à engager des efforts. Ainsi, le manque d'effort risque de détériorer la validité des inférences faites sur base des scores au test PISA en termes de compétence et limite l'usage des scores (Eklöf, 2010 ; Finn, 2015 ; Wise 2017, 2020).

Un manque d'effort altère donc l'estimation de la performance des élèves, et un niveau différent d'effort, à travers pays ou au sein d'un pays, peut impacter la comparabilité des résultats. Le présent article propose de quantifier l'effort investi par les élèves dans l'épreuve de lecture de PISA 2018. Il analyse ensuite l'effet du manque d'effort sur deux indicateurs d'efficacité et d'équité couramment rapportés, à savoir:

- l'estimation de la performance moyenne des pays, au cœur des médiatiques classements ;
- l'estimation de la différence de performance entre les filles et les garçons, thématique incontournable en sciences de l'éducation.

1.1. La compréhension de l'écrit dans PISA

La compréhension de l'écrit a fait l'objet de plusieurs études internationales, dont PISA, PIRLS (Progress in Reading Literacy Study) ou encore le PASEC (Programme d'Analyse des Systèmes Éducatifs de la Confemen) pour ne citer que les plus récentes. Ces études internationales évaluent le niveau de lecture des élèves, de différents âges ou années d'études, avec des cadres méthodologiques différents. Dans le cadre de PISA 2018, la compréhension de l'écrit se définit comme « la capacité d'un individu à comprendre, utiliser, évaluer, réfléchir sur et s'engager dans des textes afin d'atteindre des objectifs, de développer des connaissances et potentiels, et de participer à la société » (OCDE, 2019a, p. 14). La population ciblée concerne les élèves de 15 ans, soit des jeunes en fin de scolarité obligatoire dans de nombreux pays.

Les rapports PISA présentant les résultats s'ouvrent, dans un premier temps, sur les classements des pays en fonction de leur performance moyenne. Plus spécifiquement, l'OCDE distingue les pays dont la performance est statistiquement supérieure à la moyenne OCDE, dans la moyenne OCDE ou inférieure à la moyenne OCDE. Ainsi, en 2018, la moyenne de l'OCDE en compréhension de l'écrit était de 487 (OCDE, 2019b) :18 pays ont une performance moyenne supérieure à cette moyenne (les plus performants étant l'Estonie, le Canada, la Finlande, l'Irlande et la Corée), 5 pays ont une moyenne ne différant pas de la moyenne de l'OCDE et 13 pays ont une moyenne inférieure à la moyenne (les plus faibles étant la République Slovaque, la Grèce, le Chili, le Mexique et la Colombie) (OCDE, 2019b). Ces classements sont fréquemment repris dans la presse, avec un degré variable d'approfondissement (Pons, 2010), même si l'OCDE et les chercheurs promeuvent une lecture approfondie des résultats (Cattonar & Mangez, 2014 ; Champollion & Barthes, 2012 ; Grey & Morris, 2018). Cette médiatisation des classements est liée notamment à leur caractère intelligible, leur lecture à l'aune de formulations accrocheuses (Cattonar & Mangez, 2014), et s'inscrit dans une continuité des notes chiffrées traditionnellement pratiquées à l'école (Champollion & Barthes, 2012). Cependant, ces différences de rangs entre pays tendent à rassembler les pays autour de scores ne différant pas statistiquement (Champollion & Barthes, 2012 ; Duru-Bellat, 2019).

La performance en lecture est aussi mise en relation avec les caractéristiques des élèves, des écoles et des systèmes d'enseignement afin de comprendre les mécanismes sous-tendant les écarts de performance entre pays et au sein des pays. Un des prédicteurs le plus couramment étudié concerne le genre. Les différences observées en lecture tendent à être plus marquées et consistantes qu'en mathématiques et en sciences (Baye & Monseur, 2016). Ainsi, les écarts de performances moyennes entre les filles et les garçons sont en faveur des premières (Baye & Monseur, 2016 ; Mullis et al., 2017 ; OCDE, 2019c) et semblent s'accroitre avec l'âge/le grade (Baye & Monseur, 2016 ; Petersen, 2018 ; Reilly et al., 2019). Par ailleurs, la variance des performances des garçons tend à être plus élevée que celle des filles (Baye & Monseur, 2016), les garçons étant surreprésentés parmi les élèves peu performants en lecture (OCDE, 2019c ; Reilly et al., 2019).

L'ensemble de ces estimations de la performance se base sur les réponses fournies aux items du test de lecture. L'hypothèse est faite qu'elles sont le fruit du niveau de maitrise de la langue du répondant. Néanmoins, l'élève doit également avoir la volonté d'engager des efforts dans le test. Ainsi, répondre à ces items demande à la fois de l'habileté en lecture et de l'engagement dans le test (Eklöf, 2010).

1.2. L'effort investi dans le test

L'engagement dans la passation d'un test peut se définir comme « le degré auquel l'individu consacre l'effort nécessaire afin de pleinement refléter ses connaissances, compétences, et aptitudes » dans le domaine que le test vise à investiguer (Wise, 2020, p. 329). Il correspond à la motivation d'un répondant à répondre du mieux possible au test et est donc une forme spécifique de motivation d'accomplissement (Eklöf, 2010).

Dans un contexte de test à faibles enjeux, l'absence de conséquence associée à la performance à l'épreuve peut diminuer la valeur accordée par les répondants à leur performance tandis que la situation de test peut engendrer des désagréments (tels que l'anxiété, la frustration ou un sentiment de perte de temps). En accord avec l'*expectancy-value theory* (Wigfield & Eccles, 1992, 2000), cette faible valeur perçue et ces désagréments éventuels peuvent amoindrir la motivation à passer le test et résulter en de faibles efforts, se traduisant par des omissions, des réponses brèves ou au hasard (Wise & Gao, 2017). Les réponses résultant de peu d'efforts sont moins correctes que celles issues d'un comportement engagé dans le test (Wise, 2017) et les répondants engagés ont de meilleurs scores au test (Silm et al. 2020). Braun et al. (2011) ont ainsi montré que la performance des élèves à une épreuve externe variait selon les enjeux perçus par les élèves. L'enjeu est alors de déterminer dans quelle mesure une différence de performance entre élèves reflète une différence de compétence ou une différence d'effort.

L'effort investi dans le test est principalement quantifié via deux types de mesure, auto-rapportées ou basées sur les temps de réponse (TR) aux items (Finn, 2015 ; Silm et al., 2020). Les mesures auto-rapportées renvoient à une quantification de la perception qu'ont les répondants de leur propre effort (Finn, 2015). Ces instruments de mesure sont souvent administrés directement après le test de performance (Silm et al., 2020) mais peuvent également apparaître avant le test, afin de mesurer notamment la motivation initiale ou projetée (Dierendonck et al., 2013, 2017 ; Fumel & Keskpaik, 2017 ; Weis et al., 2017). Un exemple est le thermomètre PISA (Kunter et al., 2002), administré à la fin du test PISA. Les élèves doivent, dans un premier temps, imaginer une situation dans laquelle ils donneraient le meilleur d'eux-mêmes ; ils doivent ensuite se positionner, sur une échelle de 1 à 10, quant à l'effort investi dans PISA (« Par rapport à la situation que vous venez d'imaginer, quel effort pensez-vous avoir fourni en répondant à ce test ? ») et l'effort qu'ils auraient investi si PISA était coté (« Si la note obtenue lors de ce test comptait pour votre bulletin scolaire, quel effort auriez-vous fourni ? ») (OCDE, 2019b). Ces mesures issues du thermomètre présentent l'avantage de pouvoir comparer l'effort investi avec l'effort que l'élève aurait consenti si le test avait été coté. Butler et Adams (2007) proposent ainsi un indice d'effort relatif (basé sur la différence entre l'effort investi et l'effort si coté) et distinguent différents profils motivationnels d'élèves, profils affinés ultérieurement par Keskpaik et Rocher (2015). Le thermomètre PISA a connu plusieurs adaptations dans le cadre d'évaluations nationales, en France et au Luxembourg notamment, pour quantifier l'effort investi lors de la passation d'épreuves à faibles enjeux. Ainsi, plusieurs améliorations à cet instrument ont été proposées dont : des formulations simplifiées, l'utilisation des termes « application » ou « sérieux » plutôt qu'« effort » (l'effort étant lié à la difficulté relative du test, un élève performant percevant le test comme relativement facile déclarant, dès lors, engager peu d'efforts pour compléter les items) et l'ajout d'une échelle mesurant la difficulté perçue du test (Dierendonck et al., 2013 ; Keskpaik & Rocher, 2015). Keskpaik et Rocher (2015) se sont aussi interrogés sur le nombre d'échelons à proposer ; d'une part, ils proposent une version à 4 échelons pour des élèves du primaire et, d'autre part, sur base d'entretiens collectifs réalisés auprès de collégiens, les auteurs notent que ces élèves proposent d'améliorer l'instrument en passant de 10 à 5 possibilités de réponse.

Si ces mesures auto-rapportées sont relativement aisées à mettre en œuvre et permettent d'investiguer différents aspects de la motivation tels que l'effort, l'importance du test ou l'effort que l'élève aurait fait si le test était coté (Dierendonck et al., 2017 ; Wise, 2020), elles impliquent

que l'élève soit capable de décrire ses efforts, que sa perception soit précise, et qu'il réponde honnêtement (et de manière engagée) à la question (Finn, 2015 ; Wise & Gao, 2017). Les mesures basées sur les TR, quant à elles, déterminent si la réponse fournie est le fruit d'efforts sur base du temps passé sur l'item. Un seuil est établi en deçà duquel la réponse est considérée comme rapide et incompatible avec un traitement approfondi de la question[1] (Wise & Kong, 2005). La mesure s'effectue donc sans que le répondant en ait conscience (étant donc non entachée de biais de réponse) et permet une analyse de l'effort au niveau de l'item (Kong et al., 2007 ; Silm et al., 2020 ; Wise & Gao, 2017). De plus, cette mesure reposant sur des paradata, elle est plus objective que les mesures issues d'instruments auto-rapportés dont la formulation peut être ambiguë. Ainsi, le lien entre l'effort et la performance au test est plus élevé avec les mesures basées sur les TR qu'avec les mesures auto-rapportées (Silm et al., 2020).

La littérature sur l'effort investi dans le test relève qu'il peut fluctuer en fonction des caractéristiques des items et des répondants. L'effort tend à être moindre pour les items situés plus loin dans le test (Lindner et al., 2019 ; Wise, 2006 ; Wise et al., 2009), plus longs (Wise, 2006 ; Wise et al., 2009) et comprenant plus d'options de réponse (pour les items à réponse fermée) (Wise et al., 2009). Au niveau des caractéristiques des répondants, les filles semblent plus engagées dans le test (Butler & Adams, 2007 ; DeMars et al., 2013 ; Dierendonck et al., 2016 ; Keskpaik & Rocher, 2012 ; OCDE, 2015), même si certains auteurs n'observent pas de différence d'effort selon le genre (Lindner et al., 2019 ; Wise et al., 2009). Par ailleurs, l'effort tend à diminuer auprès d'élèves plus âgés (Keskpaik & Rocher, 2012 ; Weis et al., 2017) et la relation entre performance au test et effort semble augmenter avec l'âge/le grade, suggérant que les élèves plus âgés sont plus sélectifs sur les tâches dans lesquelles ils investissent des efforts (Silm et al., 2020).

Plusieurs méthodes permettent de prendre en compte, lors de l'analyse des résultats, l'effort investi dans le test. Parmi celles-ci, le filtrage sur base de la motivation consiste à retirer les individus ayant été identifiés comme non engagés afin de ne garder que ceux ayant répondu méticuleusement et pour lesquels l'hypothèse selon laquelle les réponses reflètent le niveau de compétence est tenable (Finn, 2015). Exclure ces élèves peu engagés suppose que l'effort n'est pas corrélé à la compétence (latente et non observée) des élèves (Finn, 2015 ; Wise et al., 2009) : en d'autres termes, les faibles scores associés à peu d'effort ne reflètent pas la compétence. Si effort et compétence étaient positivement corrélés, alors le filtrage entrainerait le retrait d'élèves peu compétents et l'échantillon ne serait plus représentatif, entrainant ainsi une surestimation de la compétence moyenne de la population (Finn, 2015 ; Wise et al., 2009). Notons que la sous-estimation de la performance moyenne due au manque d'effort est moindre lorsque l'effort et la compétence sont liés. En effet, la sous-estimation de la performance d'un répondant est d'autant plus grande qu'il est compétent. La distorsion étant donc quasi nulle chez les individus peu compétents, la performance moyenne de la population est moins sous-estimée lorsque compétence et effort sont positivement corrélés que lorsqu'ils sont indépendants (Wise, 2020).

La littérature apporte des résultats inconsistants sur la relation entre effort et compétence. Certains auteurs (DeMars et al., 2013 ; Kong et al., 2007 ; Wise & Kong, 2005) ne trouvent pas de relation entre l'effort et la compétence, estimée via des mesures externes de performance. Cependant, d'autres auteurs observent une association positive entre l'effort et des mesures externes de compétence (Wise et al., 2009) ou des mesures cognitives (Lindner et al., 2019).

[1] Une réponse rapide est considérée comme désengagée dans un test de puissance (la performance étant mesurée à travers la réussite des items) et non dans un test de vitesse (où la performance est directement reflétée dans la rapidité de la réponse).

1.3. But de l'article

La performance en lecture estimée par PISA peut donc souffrir d'un manque d'effort des élèves dans le test. Sur base des données issues du cycle 2018 des pays membres de l'OCDE, le présent article analyse l'effort estimé sur base des TR aux items. Une première analyse concerne les caractéristiques des réponses rapides en vue de confirmer la validité de cette mesure d'effort dans le contexte de PISA 2018. L'effet de l'effort sur le classement international des pays (basé sur leur moyenne) est ensuite investigué par le biais de filtrage. De plus, il est attendu que les garçons dépensent moins d'effort dans le test ; leur performance en lecture est donc potentiellement plus sous-estimée que celle des filles. L'effet de ce différentiel sur les différences de genre en matière de compréhension de l'écrit est également investigué.

2. Méthodologie

2.1. Données

Le test cognitif PISA 2018 a été administré sur ordinateur dans la majorité des pays. L'épreuve se structure en deux sessions de test, avec une session vouée à la lecture. Environ la moitié des étudiants a donc été évaluée dans ce domaine lors de la première session et l'autre moitié lors de la seconde session de test.

Alors que le test en lecture était précédemment linéaire, ce domaine est évalué par le biais d'un test adaptatif séquentiel (*multistage testing*) pour la première fois en 2018. Il se déroule en trois étapes, les modules administrés à chaque étape étant constitués de plusieurs unités d'items (OCDE, n.d.). La première étape est constituée de modules de routage tandis que les étapes 2 et 3 proposent des modules relativement faciles ou difficiles[2] selon la performance antérieure de l'élève. Les modules présentés à chaque étape sont donc fonction de la performance précédemment estimée de l'élève, selon un routage probabiliste : par exemple, un étudiant réalisant une piètre performance au premier module sera probablement orienté ultérieurement vers un module facile (Mead, 2006).

L'effort investi dans le test étant mesuré par le biais des TR aux items, les langues de test sont distinguées lors de l'étape de construction de l'indice (voir point 2.2), une langue étant analysée si au minimum 3000 étudiants ont passé le test dans cette langue (tous pays et sessions confondus). La Norvège n'a pas été retenue car les données relatives à la langue de test ne sont pas disponibles. Enfin, les élèves ayant reçu une version simplifiée du test (livret Une Heure, UH) sont exclus. Les tailles d'échantillon pour les 36 pays de l'OCDE sélectionnés sont présentées dans le tableau 1.

[2] Certaines unités sont administrées dans des modules des deux niveaux de difficulté à la troisième étape.

Tableau 1. Par pays et langue de test, nombre d'étudiants exclus (autre langue, livret UH, identifiant du livret ou des modules invalide, absence de TR) et retenus pour analyse.

			n exclus				n analysés	
Pays		**Langue**	**Lang.**	**Livret UH**	**Identifiant invalide**	**Absence de TR**	**Session 1**	**Session 2**
Australie	AUS	Anglais				42	7153	7078
Autriche	AUT	Allemand		83		1	3364	3354
Belgique	BEL	Néerlandais		127		7	2365	2383
		Français		157		7	1517	1540
		Allemand		7		2	186	177
Canada	CAN	Anglais		588		6	8108	7986
		Français		217		14	2856	2878
Suisse	CHE	Français				1	706	705
		Allemand				4	1735	1716
		Italien				1	476	478
Chili	CHL	Espagnol			1	19	3816	3785
Colombie	COL	Espagnol				19	3754	3749
Rép. Tchèque	CZE	Tchèque		30		2	3469	3518
Allemagne	DEU	Allemand		98			2684	2669
Danemark	DNK	Danois		494		15	3569	3579
Espagne	ESP	Catalan	2339			2	1542	1478
		Espagnol				47	15342	15193
Estonie	EST	Estonien				2	1979	2019
		Russe				1	659	656
Finlande	FIN	Finnois		37		3	2647	2576
		Suédois		4			186	196
France	FRA	Français				7	3121	3180
Royaume Uni	GBR	Anglais	458			33	6678	6649
Grèce	GRC	Grec				6	3179	3218
Hongrie	HUN	Hongrois				7	2552	2573
Irlande	IRL	Anglais	76			14	2731	2756
Islande	ISL	Islandais		66		2	1609	1619
Israël	ISR	Hébreu	1658		756	3	2098	2108
Italie	ITA	Allemand					554	559
		Italien			2	5	5309	5356
Japon	JPN	Japonais				2	3059	3048
Corée	KOR	Coréen				3	3321	3326
Lituanie	LTU	Lituanien				3	2951	2914
		Polonais					237	238
		Russe					274	268
Luxembourg	LUX	Anglais					125	137
		Français				1	809	823
		Allemand					1670	1665
Lettonie	LVA	Letton				3	1895	1961
		Russe					722	722
Mexique	MEX	Espagnol				9	3614	3676
Pays-Bas	NLD	Néerlandais		851		5	1951	1958
N. Zélande	NZL	Anglais				5	3091	3077
Pologne	POL	Polonais				1	2785	2839
Portugal	PRT	Portugais			1	4	2983	2944
Slovaquie	SVK	Hongrois				2	142	148
		Slovaque		148		9	2755	2761
Slovénie	SVN	Slovène		160		4	3125	3112
Suède	SWE	Anglais					9	11
		Suédois			3	12	2707	2762
Turquie	TUR	Turc				5	3435	3450
USA	USA	Anglais		39		17	2380	2402

2.2. Mesures d'effort

La principale mesure d'effort analysée se base sur les TR. Pour chaque distribution de TR d'item, par session et par langue de test, les réponses rapides (et probablement résultant de peu d'effort) sont distinguées du reste de la distribution ; le seuil en deçà duquel une réponse est définie comme rapide est fixé selon la méthode du *Normative Threshold 10* (NT10 : Wise & Gao, 2017). Ainsi, une réponse est rapide si son TR est inférieur à un dixième du temps moyen de réponse, avec un seuil maximal de 10 secondes. Pour chacun des 239 TR d'item, ce sont donc 52 distributions (2 sessions * 26 langues) qui sont analysées. Deux seuils pour chaque item dans chaque langue sont ainsi calculés afin d'identifier au mieux le désengagement dans chaque session. Cet article analysant séparément les sessions, ce procédé permet d'accroitre la validité des mesures d'effort ; entre sessions, les indices d'effort ne sont cependant pas comparables, limitant l'interprétation des résultats. Si le TR_{ij} de l'élève j pour l'item i est inférieur au seuil T_i de sa distribution, la réponse est désengagée et l'indice de recherche de solution (*Solution Behavior*, SB) SB_{ij} est égal à 0 ; inversement, si le TR_{ij} est supérieur à T_i, la réponse est engagée et SB_{ij} vaut 1 (Wise & Kong, 2005, p. 167) :

$$SB_{ij} = \begin{cases} 1 \text{ si } TR_{ij} > T_i \\ 0 \text{ si } TR_{ij} \leq T_i \end{cases} \quad (1)$$

Ces indices sont ensuite agrégés au niveau élève dans un indice RTE (*Response-Time Effort*), l'indice RTE_j de l'élève j étant la moyenne des k indices SB (Wise & Kong, 2005, p. 168) :

$$RTE_j = \frac{\sum_{i=1}^{k} SB_{ij}}{k} \quad (2)$$

Enfin, les mesures auto-rapportées issues du thermomètre PISA sont également analysées : sur une échelle de 1 à 10, « […] quel effort pensez-vous avoir fourni en répondant à ce test ? » (effort investi) et « si la note obtenue lors de ce test comptait pour votre bulletin scolaire, quel effort auriez-vous fourni ? » (effort si coté).

2.3. Analyses

Les analyses se déroulent en 2 temps. Une première approche descriptive de l'effort investi dans les items, langue par langue, analyse la validité des mesures d'effort basées sur les TR. Dans un second temps, au niveau pays, des analyses inférentielles relatives à l'estimation de l'effort investi dans le test et son effet sur l'estimation de la performance sont présentées.

2.3.1. Description des réponses rapides

Les caractéristiques des réponses rapides sont décrites en termes de pourcentages d'omission et de réussite[3], les réponses rapides étant supposées être moins correctes et plus fréquemment omises. Pour cinq TR, deux items sont associés au TR (par exemple, 2 questions sur un même écran) et seule la réponse au premier item est analysée. Ces analyses étant purement descriptives, elles ne sont pas pondérées. Ensuite, au niveau des items, les taux moyens de désengagement (soit 1-*SB*) selon 1) l'étape du module, 2) sa difficulté (étapes 2 et 3) et 3) le format d'item sont détaillés. Il est attendu que l'effort diminue à travers les étapes et qu'il est également plus faible aux items ouverts.

[3] Les crédits partiels sont codés comme incorrects et les omissions non prises en compte dans l'analyse. Si un item a 0 élève désengagé (ou 0 réponse non omise parmi les réponses désengagées) dans une combinaison langue*session, il n'est pas repris dans l'analyse de cette combinaison.

2.3.2. Quantification de l'effet de l'effort sur les résultats

Par pays, l'effort moyen des élèves (indice RTE) et les corrélations entre cet indice et les deux indices d'effort rapportés dans le thermomètre PISA sont présentés. De plus, l'indice RTE moyen et la performance moyenne des élèves sont calculés en fonction des profils motivationnels des élèves, basés sur leurs réponses au thermomètre, dans la continuité de Kespkaik et Rocher (2015):

- les élèves « irréalistes », dont l'effort investi est supérieur à l'effort si coté. Contrairement à ces élèves, les 6 profils suivants ont tous indiqué un effort investi inférieur ou égal à l'effort si coté ;
- les élèves « démotivés », dont les efforts investis et si cotés sont inférieurs à 8 ;
- les élèves « pragmatiques », dont l'effort si coté est élevé (≥8) et l'effort investi faible (≤5) ;
- les élèves « peu motivés », dont l'effort si coté est élevé (≥8) et l'effort investi plus faible (6, ou 7 uniquement si l'effort si coté est de 10) ;
- les élèves « réalistes », dont l'effort si coté est élevé (≥9) et l'effort investi plus faible de 2 points ;
- les élèves « assidus », dont l'effort si coté est élevé (≥8) et l'effort investi plus faible d'un point ;
- les élèves « partisans », dont l'effort si coté et l'effort investi sont élevés (≥8) et égaux.

Ces premières analyses visent à analyser, au regard des mesures auto-rapportées, la validité de l'indice RTE pour l'estimation de la motivation à passer le test. Ensuite, la performance moyenne en lecture et les différences de genre en la matière sont estimées, par session de test, avant et après filtrage. Ce dernier consiste à recalculer les statistiques nationales après exclusion des élèves désengagés, c'est-à-dire dont le RTE est inférieur à 0,75 : ces élèves ont donc répondu rapidement à au moins un quart de leur test. Trois indices de différences filles-garçons sont présentés : les différences de moyennes, les ampleurs de l'effet (des différences égales à 0 traduisant une absence d'effet) (équation 4) et les ratios de variance (un ratio égal à 1 signifiant une variance égale) (équation 3) (Baye & Monseur, 2016) :

$$\frac{\hat{\sigma}^2_{garçons}}{\hat{\sigma}^2_{filles}} \tag{3}$$

$$\frac{\hat{\mu}_{garçons} - \hat{\mu}_{filles}}{\sqrt{\frac{\hat{\sigma}^2_{garçons} + \hat{\sigma}^2_{filles}}{2}}} \tag{4}$$

Ces analyses sont pondérées et en adéquation avec la méthodologie des valeurs plausibles, et les erreurs types sont obtenues par réplication (OCDE, 2009) ; le seuil d'erreur de première espèce est fixé à 0,05. En regard des estimations nationales, la moyenne arithmétique de l'OCDE est également présentée.

3. Résultats

3.1. Description des réponses rapides aux items

L'indice d'effort aux items a été estimé, par langue et session, en utilisant le seuil NT10. Le point 3.1.1 s'ouvre sur une illustration d'une réponse rapide et décrit les caractéristiques de ces réponses rapides. Ensuite, dans le point 3.1.2, les caractéristiques des items en termes de pourcentage de réponses rapides sont analysées.

3.1.1. Caractéristiques des réponses rapides

À titre d'exemple, le premier item de l'unité Rapa Nui (R551Q01) est présenté à la figure 1, dans sa version francophone. Pour cet item, les TR moyens sont de 88,9s (pour la session 1) et 86,3s (session 2) ; les seuils en deçà desquels les réponses sont qualifiées de rapides sont donc de 8,9s et 8,6s, respectivement (pour rappel, une réponse est rapide si elle a été formulée en moins d'un dixième du TR moyen, avec un maximum de 10 secondes). Un TR inférieur à 9s est peu compatible avec un traitement approfondi de la question et du texte ; une telle réponse est donc susceptible de relever du hasard.

Ces réponses rapides sont caractérisées par plus d'omissions et moins de réponses correctes que les réponses non rapides. La différence de pourcentage moyen de réussite (par langue et session) entre les réponses rapides et non rapides varie entre -34% (en français, pour les 2 sessions) et -51% (hébreu, session 1) et -47% (coréen, session 2) : pour le test francophone, le pourcentage de réussite est, en moyenne, moindre de 34% pour les réponses rapides. Pareillement, la moyenne des différences de pourcentage d'omission varie entre 38% (estonien) et 54% (espagnol) pour la session 1 et entre 37% (letton) et 55% (suédois) pour la session 2 ; en français, ces différences sont de l'ordre de 46% (session 1) et 49% (session 2). Les réponses rapides sont donc plus souvent des omissions et, lorsqu'une réponse est fournie, celle-ci est plus fréquemment erronée.

Figure 1. Item 1 de l'unité Rapa Nui.

Note. Extrait de : https://pisa2018-questions.oecd.org/platform/index.html?user=&domain=REA&unit=R551-RapaNui&lang=fra-ZZZ

3.1.2. Caractéristiques des items qui reçoivent moins d'effort

Le pourcentage de réponses rapides aux items est analysé au regard des caractéristiques de ceux-ci. Concernant les formats d'items, ceux à réponse ouverte sont moins susceptibles de recevoir des efforts que les items à choix multiples. En moyenne, à travers les différentes langues de test retenues dans le cadre de cette étude, le pourcentage moyen de désengagement des items ouverts est plus élevé que celui des items à réponse fermée de 1,67% pour la session 1 et de 2,44% pour la session 2. En session 1, le taux brut de réponses rapides est de 3,4% (session 2 : 4%) pour les items à choix multiples et de 5,07% (session 2 : 6,44%) pour ceux à réponse ouverte. Ces différences selon le format varient entre langues de test : elles sont faibles en néerlandais (en moyenne, 0,02% et 0,12% de différence, pour la session 1 et 2) et en estonien (0,59% et 0,56%, respectivement) et élevées en grec (3,51% et 4,82%).

Le taux de réponses rapides fluctue également selon les caractéristiques des modules (étape et difficulté) dans lesquels ils sont administrés ; les moyennes de ces pourcentages moyens, à travers les langues de tests, sont présentées dans le tableau 2. Le taux moyen de désengagement aux items augmente à travers les étapes. À travers les différentes langues, la moyenne est, pour la session 1, de 1,25% (2,84% en session 2) pour l'étape 1, 2,03% (3,24%) pour l'étape 2 et 7,12% (7,15%) pour l'étape 3. Par ailleurs, le niveau de difficulté des modules est lié à l'effort investi dans les items à l'étape 2 (pour rappel, l'étape 1 est constituée de modules de routage) : les pourcentages moyens de réponses rapides aux items sont, en moyenne, de 1,13% (session 1) et 1,69% (session 2) pour les modules difficiles et de, respectivement, 2,83% et 4,67% pour les modules faciles. Ainsi, les élèves désengagés dans le test lors du module de routage réalisent une piètre performance et sont orientés subséquemment vers des modules faciles. Dans la dernière étape, en revanche, aucun pattern ne se dégage, possiblement car le plus grand désengagement observé dans les modules faciles est contrebalancé par de la fatigue et du manque de temps (engendrant des réponses rapides) dans les modules difficiles.

Tableau 2. Moyenne des pourcentages moyens (par langue de test) de réponses rapides aux items, par session de test et par étape. Pour les étapes 2 et 3, les niveaux de difficulté des modules dans lesquels les items ont été administrés sont distingués. Les élèves ayant reçu le design alternatif avec les chemins étape 1 → étape 3 → étape 2 (approximativement 25% des étudiants) ont été exclus de cette analyse.

		Étape 2			**Étape 3**			
Session de test	Étape 1	Tous les modules	Modules difficiles	Modules faciles	Tous les modules	Modules difficiles	Modules faciles & difficiles	Modules faciles
1	1,25	2,03	1,13	2,83	7,12	7,48	7,05	6,88
2	2,84	3,24	1,69	4,67	7,15	6,08	7,45	7,74

3.2. Description de l'effort investi par l'élève

Alors que le point 3.1 analysait l'effort en fonction de certaines caractéristiques des items, la suite de ce texte propose une analyse de l'effort au niveau élève (indice *RTE*). Les moyennes nationales de cet indice sont présentées dans le tableau 3, toujours en distinguant les deux sessions de test.

Pour la première session, l'effort moyen dans le test varie entre 0,94 (Grèce) et 0,99 (Irlande), la moyenne arithmétique étant de 0,97 : en moyenne, dans l'OCDE, les élèves répondent avec effort à 97% des items qui leur sont présentés. Pour la seconde session, la moyenne OCDE de l'effort est égale à 0,96.

Par ailleurs, l'effort dans le test inféré au départ des TR tend à corréler, mais très faiblement, avec l'effort auto-rapporté et mesuré par le thermomètre PISA (indice d'effort investi), comme indiqué

dans le tableau 4. Pour la session 1, cette corrélation ne diffère pas statistiquement de 0 dans 4 pays (Colombie, France, Japon et Turquie) ; à l'inverse, elle est supérieure à 0,2 en Australie, en Finlande, en Islande, en Corée, aux Pays-Bas et aux USA. La corrélation moyenne est de 0,13. Pour la seconde session, la corrélation entre les deux mesures d'effort diffère statistiquement de 0 dans tous les pays à l'exception de la Colombie et est supérieure à 0,2 dans 18 pays, la corrélation moyenne dans l'OCDE étant de 0,19. Les corrélations entre l'indice RTE et l'effort que les étudiants auraient investi si PISA avait compté pour le bulletin (effort si coté) sont également présentées. Ces corrélations sont relativement similaires, en moyenne, à celles observées avec l'indice d'effort investi, même si des différences existent entre pays. Ainsi, les différences entre les corrélations avec l'indice d'effort investi et avec l'effort si coté sont particulièrement élevées dans les pays présentant une corrélation élevée avec l'effort investi. Par exemple, aux Pays-Bas, la corrélation avec l'indice d'effort investi est de 0,31 en session 1 et de 0,35 en session 2 ; ces corrélations, une fois calculées avec l'indice d'effort si coté, chutent à 0,16 et 0,21, respectivement, soit des valeurs plus proches de la moyenne OCDE. De même, il y a moins de variabilité dans les corrélations du RTE avec l'indice d'effort si coté que dans celles avec l'effort investi.

L'évolution de l'indice RTE moyen en fonction des profils motivationnels est ensuite analysée. La figure 2 présente la moyenne OCDE du RTE moyen et de la performance moyenne en lecture par profil motivationnel ; le tableau 5 présente les valeurs associées à cette figure. Les profils les plus fréquents sont, en moyenne, les élèves assidus (22%), les partisans (21% session 1 et 22% session 2) et les réalistes (21% et 20% respectivement) tandis que les élèves peu motivés représentent, en moyenne, environ 15% des élèves. Les élèves réalistes et assidus sont ceux qui, en moyenne, ont engagé le plus d'effort dans le test (moyenne OCDE du RTE moyen de 0,98) et ont obtenu les meilleures performances en lecture ; les élèves peu motivés et partisans ont réalisé légèrement moins d'effort (en moyenne, 0,97 pour la session 1 et 0,96 pour la session 2) et ont des performances en lecture légèrement inférieures à celles des réalistes et des assidus. Enfin, les élèves pragmatiques (8%), démotivés et irréalistes (6%) sont moins fréquents ; ces élèves ont engagé en moyenne moins d'effort que les élèves peu motivés, réalistes, assidus et partisans, et leur performance est également nettement plus faible.

Ces résultats soulignent que le thermomètre PISA, qui est administré en fin de test, fournit une mesure d'effort relativement consistante avec l'effort engagé dans la seconde session de test ; dans la première session, le lien est très ténu, probablement en raison de l'administration de la seconde partie du test entre ces deux mesures. L'indice RTE moyen est également consistant avec les profils motivationnels des élèves, basés sur les données auto-rapportées du thermomètre. Ainsi, les élèves ayant soit déclaré peu d'effort sur les deux dimensions du thermomètre PISA (démotivés), soit un effort si coté élevé et un effort investi faible (pragmatiques), soit une réponse irréaliste (indiquant un effort investi plus élevé que si coté), sont effectivement les profils ayant réalisé, en moyenne, le moins d'effort et le plus de réponses rapides. Comme attendu au regard des plus faibles probabilités de succès de ces réponses rapides, ces élèves ont des performances moyennes plus faibles.

Étant donné qu'un manque d'effort tend à être associé à une sous-estimation de la performance, les prochains points détaillent son effet sur l'estimation (a) de la performance moyenne et (b) des différences de performances entre filles et garçons. Ces estimations sont calculées, d'une part, sur l'échantillon complet et, d'autre part, après exclusion des élèves ayant fourni peu d'efforts (pour rappel, au moins 25% des réponses rapides). Le nombre d'élèves exclus, par session, et le pourcentage de filles exclues sont présentés dans le tableau 3. En moyenne, 3% des élèves sont exclus dans la session 1 et 5% dans la session 2. Ces élèves sont majoritairement des garçons (les seuls pays pour lesquels le pourcentage de filles exclues excède 50% étant le Chili et la Colombie, en session 1).

Tableau 3. Par session, effort moyen estimé par l'indice RTE ($\hat{\mu}$), nombre élèves ayant été évalués en lecture et nombre d'élèves exclus (dont le pourcentage de filles)

	Session 1				Session 2			
Pays	$\hat{\mu}$	SE	n	n filtré (% filles)	$\hat{\mu}$	SE	n	n filtré (% filles)
AUS	0,97	0,00	7153	245 (37%)	0,96	0,00	7078	376 (36%)
AUT	0,97	0,00	3364	95 (38%)	0,96	0,00	3354	141 (26%)
BEL	0,98	0,00	4068	77 (40%)	0,97	0,00	4100	134 (34%)
CAN	0,98	0,00	10964	288 (33%)	0,97	0,00	10864	355 (35%)
CHE	0,97	0,00	2917	82 (29%)	0,95	0,00	2899	154 (36%)
CHL	0,95	0,00	3816	235 (52%)	0,95	0,00	3785	204 (45%)
COL	0,93	0,01	3754	414 (54%)	0,95	0,00	3749	183 (49%)
CZE	0,97	0,00	3469	75 (37%)	0,95	0,00	3518	178 (40%)
DEU	0,97	0,00	2684	80 (35%)	0,96	0,00	2669	86 (27%)
DNK	0,97	0,00	3569	99 (40%)	0,97	0,00	3579	149 (34%)
ESP	0,97	0,00	16884	394 (40%)	0,96	0,00	16671	597 (33%)
EST	0,98	0,00	2638	50 (40%)	0,98	0,00	2675	61 (36%)
FIN	0,98	0,00	2833	54 (17%)	0,97	0,00	2772	90 (22%)
FRA	0,96	0,00	3121	158 (45%)	0,95	0,00	3180	185 (36%)
GBR	0,95	0,00	6678	312 (37%)	0,95	0,00	6649	331 (33%)
GRC	0,94	0,00	3179	225 (46%)	0,93	0,00	3218	272 (33%)
HUN	0,97	0,00	2552	68 (38%)	0,96	0,00	2573	99 (29%)
IRL	0,99	0,00	2731	32 (34%)	0,98	0,00	2756	56 (14%)
ISL	0,97	0,00	1609	68 (35%)	0,95	0,00	1619	108 (39%)
ISR	0,96	0,00	2098	106 (27%)	0,93	0,01	2108	171 (31%)
ITA	0,95	0,00	5863	258 (42%)	0,94	0,00	5915	388 (29%)
JPN	0,97	0,00	3059	81 (32%)	0,97	0,00	3048	95 (38%)
KOR	0,97	0,00	3321	115 (35%)	0,95	0,00	3326	208 (36%)
LTU	0,98	0,00	3462	70 (27%)	0,96	0,00	3420	125 (18%)
LUX	0,97	0,00	2604	71 (18%)	0,95	0,00	2625	133 (29%)
LVA	0,97	0,00	2617	58 (31%)	0,96	0,00	2683	110 (35%)
MEX	0,98	0,00	3614	59 (49%)	0,98	0,00	3676	49 (45%)
NLD	0,97	0,00	1951	58 (21%)	0,96	0,00	1958	93 (38%)
NZL	0,97	0,00	3091	89 (42%)	0,96	0,00	3077	135 (28%)
POL	0,97	0,00	2785	65 (28%)	0,96	0,00	2839	103 (27%)
PRT	0,97	0,00	2983	105 (32%)	0,96	0,00	2944	129 (39%)
SVK	0,95	0,00	2897	136 (49%)	0,94	0,00	2909	189 (40%)
SVN	0,98	0,00	3125	42 (21%)	0,97	0,00	3112	123 (20%)
SWE	0,95	0,00	2716	186 (47%)	0,93	0,01	2773	255 (40%)
TUR	0,98	0,00	3435	49 (31%)	0,98	0,00	3450	40 (23%)
USA	0,98	0,00	2380	38 (24%)	0,97	0,00	2402	68 (25%)
Moy.	0,97			3% (36%)	0,96			5% (33%)

Tableau 4. Corrélation ($\hat{\rho}$) de l'effort RTE avec les indices auto-rapportés (i) d'effort investi dans PISA (« Par rapport à la situation que vous venez d'imaginer, quel effort pensez-vous avoir fourni en répondant à ce test ? ») et (ii) d'effort si PISA avait compté pour des points (« Si la note obtenue lors de ce test comptait pour votre bulletin scolaire, quel effort auriez-vous fourni ? »)

	Session 1					**Session 2**				
		Effort investi		Effort si coté			Effort investi		Effort si coté	
CNT	N suppr.	$\hat{\rho}$	SE	$\hat{\rho}$	SE	N suppr.	$\hat{\rho}$	SE	$\hat{\rho}$	SE
AUS	1083(15%)	0,20	0,02	0,18	0,03	1115(16%)	0,26	0,02	0,21	0,02
AUT	313(9%)	0,17	0,04	0,10	0,03	330(10%)	0,21	0,03	0,13	0,03
BEL	442(11%)	0,06	0,02	0,06	0,03	469(11%)	0,17	0,03	0,16	0,02
CAN	1434(13%)	0,14	0,02	0,13	0,02	1479(14%)	0,22	0,02	0,15	0,02
CHE	313(11%)	0,18	0,03	0,13	0,04	330(11%)	0,23	0,03	0,17	0,03
CHL	524(14%)	0,07	0,03	0,07	0,04	646(17%)	0,11	0,03	0,13	0,03
COL	794(21%)	0,02	0,02	0,06	0,02	822(22%)	0,05	0,03	0,06	0,02
CZE	247(7%)	0,14	0,03	0,12	0,03	272(8%)	0,13	0,03	0,15	0,03
DEU	286(11%)	0,11	0,03	0,17	0,05	279(10%)	0,20	0,03	0,23	0,03
DNK	317(9%)	0,20	0,03	0,15	0,03	345(10%)	0,23	0,04	0,16	0,03
ESP	2594(15%)	0,11	0,02	0,12	0,02	3244(19%)	0,21	0,02	0,18	0,02
EST	199(8%)	0,08	0,02	0,01	0,02	222(8%)	0,11	0,03	0,10	0,04
FIN	301(11%)	0,22	0,03	0,23	0,04	317(11%)	0,38	0,04	0,28	0,04
FRA	475(15%)	0,05	0,03	0,07	0,03	527(17%)	0,10	0,02	0,11	0,02
GBR	860(13%)	0,14	0,03	0,17	0,03	950(14%)	0,25	0,03	0,20	0,03
GRC	349(11%)	0,09	0,02	0,11	0,03	332(10%)	0,09	0,02	0,12	0,03
HUN	225(9%)	0,06	0,03	0,13	0,03	242(9%)	0,16	0,03	0,18	0,03
IRL	320(12%)	0,14	0,03	0,14	0,04	309(11%)	0,16	0,03	0,10	0,03
ISL	162(10%)	0,23	0,04	0,22	0,05	157(10%)	0,40	0,03	0,28	0,04
ISR	187(9%)	0,17	0,04	0,15	0,05	183(9%)	0,29	0,04	0,16	0,03
ITA	695(12%)	0,13	0,03	0,09	0,03	756(13%)	0,25	0,04	0,27	0,05
JPN	221(7%)	0,05	0,03	0,10	0,04	230(8%)	0,08	0,02	0,16	0,03
KOR	278(8%)	0,21	0,04	0,20	0,03	251(8%)	0,26	0,04	0,23	0,04
LTU	306(9%)	0,11	0,03	0,17	0,04	346(10%)	0,17	0,03	0,15	0,03
LUX	314(12%)	0,19	0,03	0,19	0,04	310(12%)	0,20	0,03	0,21	0,03
LVA	220(8%)	0,13	0,03	0,17	0,02	268(10%)	0,16	0,03	0,23	0,03
MEX	344(10%)	0,07	0,03	0,07	0,02	390(11%)	0,06	0,02	0,07	0,03
NLD	191(10%)	0,31	0,04	0,16	0,04	216(11%)	0,35	0,04	0,21	0,06
NZL	424(14%)	0,14	0,03	0,12	0,02	522(17%)	0,21	0,03	0,21	0,03
POL	314(11%)	0,08	0,02	0,08	0,02	357(13%)	0,17	0,02	0,16	0,02
PRT	320(11%)	0,08	0,02	0,10	0,03	371(13%)	0,14	0,04	0,19	0,03
SVK	321(11%)	0,05	0,02	0,07	0,02	318(11%)	0,15	0,03	0,19	0,03
SVN	159(5%)	0,07	0,03	0,14	0,04	170(5%)	0,22	0,03	0,18	0,03
SWE	293(11%)	0,09	0,03	0,14	0,03	327(12%)	0,21	0,03	0,19	0,04
TUR	323(9%)	0,04	0,03	0,07	0,03	369(11%)	0,06	0,02	0,04	0,02
USA	328(14%)	0,22	0,03	0,13	0,04	361(15%)	0,19	0,04	0,21	0,04
Moy.	11%	0,13		0,13		12%	0,19		0,17	

Note. Le nombre d'étudiants exclus présenté dans la première colonne pour les résultats de chaque session correspond au nombre/pourcentage d'élèves n'ayant pas répondu à au moins un des deux indicateurs d'effort auto-rapporté.

Tableau 5. Moyenne OCDE du RTE moyen, de la performance moyenne en lecture et de la fréquence par profil motivationnel des élèves

		Profil motivationnel des élèves						
Session	Statistique	Irréalistes	Démotivés	Pragmatiques	Peu motivés	Réalistes	Assidus	Partisans
1	Fréq. moyenne	6,00	6,41	8,08	15,45	20,76	22,29	21,01
	$\hat{\mu}$ RTE moyenne	0,95	0,94	0,95	0,97	0,98	0,98	0,97
	$\hat{\mu}$ perf. moyenne lecture	430,11	454,99	470,63	503,32	519,68	526,91	505,30
2	Fréq. moyenne	6,14	6,16	7,92	15,09	20,33	21,96	22,40
	$\hat{\mu}$ RTE moyenne	0,93	0,91	0,92	0,96	0,98	0,98	0,96
	$\hat{\mu}$ perf. moyenne lecture	413,18	429,24	440,07	481,16	502,00	511,35	488,50

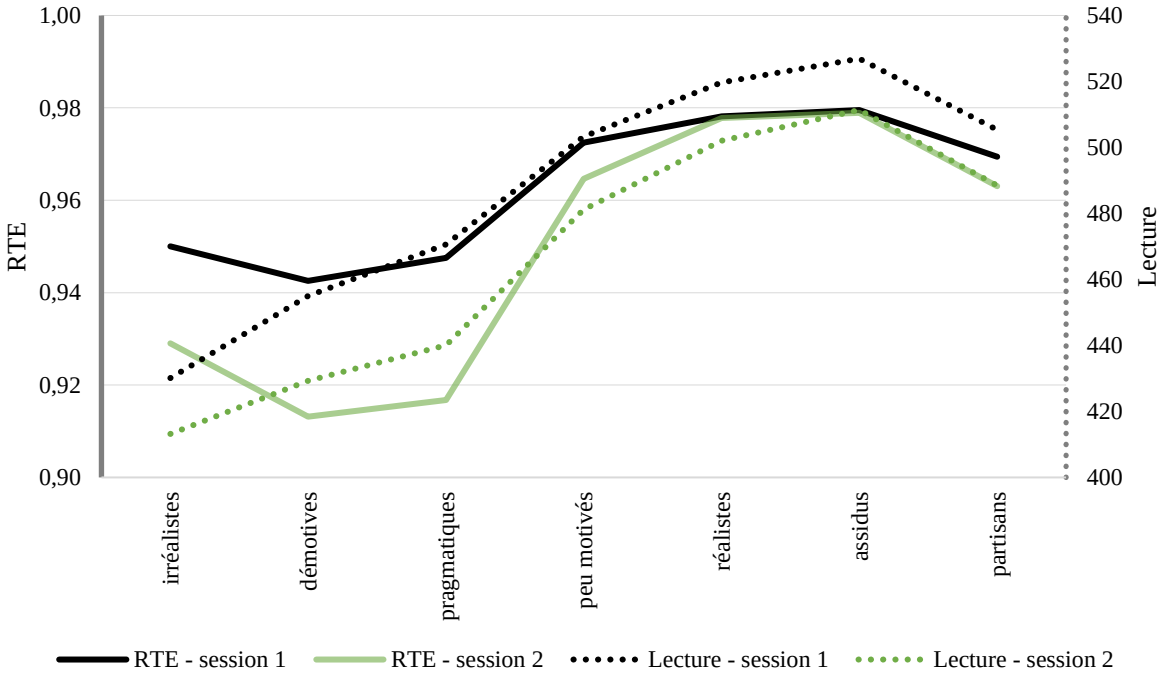

Figure 2. Moyenne OCDE de la performance moyenne en lecture (en pointillé) et de l'indice RTE moyen (en trait continu), par profil motivationnel

3.3. Effort et performances moyennes

Le tableau 6 présente, par session, l'estimation de la performance moyenne en lecture (et le classement associé) avant et après filtrage des élèves peu engagés dans le test, ainsi que la différence de moyennes et de rangs. Comme attendu, les moyennes après filtrage tendent à augmenter. Ainsi, la performance moyenne des pays de l'OCDE passe de 498 à 502 pour la première session de test et de 479 à 486 pour la seconde session. L'effet du filtrage sur la moyenne varie entre pays, la moyenne de certains pays étant quasiment inchangée (comme au Mexique ou en Estonie lors de la première session) alors que d'autres pays ont une forte augmentation de leur moyenne (tel Israël, dont la moyenne passe, lors de la session 2, de 487 à 503 après filtrage). Ces différences de moyennes sont liées à l'ampleur de l'effort investi dans le test : elles corrèlent à -0,33 (session 1) et -0,79 (session 2) avec l'effort moyen investi dans la session de test (indice *RTE*, tableau 3), les pays dans lesquels l'effort est le plus élevé ayant une différence de score moindre.

Ces modifications peuvent impacter substantiellement le classement de certains pays ; par exemple, en session 2, le « top 3 » est composé de l'Estonie, la Finlande et l'Irlande avant filtrage et de la Finlande, l'Estonie et la Corée après filtrage (l'Irlande étant désormais 5^{e}). Ces modifications de places sont importantes lors de la seconde session de test (l'effort y étant moindre), la plus forte variation étant observée pour Israël, qui passe de la 17^{e} place à la 9^{e} place. Il est à noter que de plus faibles modifications de classement sont observées parmi les pays les moins performants, probablement en raison des plus importants écarts de moyennes entre eux.

Si le classement varie substantiellement après filtrage, ce n'est cependant pas le cas de la significativité des différences entre les moyennes nationales et la moyenne OCDE. Ainsi, pour la première session, le seul changement après filtrage concerne la France, dont la moyenne nationale devient statistiquement significative (supérieure à la moyenne OCDE). Pour la seconde session, après filtrage, la moyenne d'Israël devient supérieure à la moyenne OCDE et celle de la Suisse statistiquement non différente de cette dernière. Dans tous les autre cas, l'interprétation de la moyenne nationale au regard de la moyenne OCDE reste inchangée après filtrage.

Tableau 6. Par session, moyenne estimée en lecture et rang (Rg., en ordre descendant) de la performance en lecture, avant et après filtrage. Les différences de moyennes et de rangs (après-avant) sont présentées

	Session 1								Session 2							
	Avant filtrage			Après filtrage			Différence			Avant filtrage			Après filtrage			Différence	
Pays	$\hat{\mu}$	SE	Rg.	$\hat{\mu}$	SE	Rg.	$\hat{\mu}$	Rg.	Pays	$\hat{\mu}$	SE	Rg.	$\hat{\mu}$	SE	Rg.	$\hat{\mu}$	Rg.
EST	532	2	1	533	2	2	1	1	EST	515	2	1	518	2	2	3	1
CAN	532	2	2	534	2	1	3	-1	FIN	514	3	2	520	3	1	6	-1
FIN	529	2	3	532	2	3	3	0	IRL	513	3	3	516	3	5	3	2
KOR	523	3	4	530	3	5	7	1	CAN	512	2	4	516	2	4	4	0
IRL	523	3	5	523	3	6	1	1	KOR	505	4	5	518	3	3	13	-2
ISR	522	5	6	531	4	4	10	-2	POL	505	3	6	511	3	6	6	0
POL	519	3	7	522	3	7	3	0	JPN	497	3	7	502	3	10	5	3
NZL	517	2	8	521	2	8	4	0	USA	497	4	8	501	4	13	4	5
SWE	517	4	9	521	3	9	4	0	DNK	496	2	9	500	2	14	5	5
USA	515	4	10	518	4	12	3	2	SWE	496	3	10	506	3	7	11	-3
GBR	514	3	11	520	2	10	6	-1	GBR	495	3	11	503	3	8	8	-3
DNK	514	2	12	517	2	13	3	1	NZL	494	3	12	502	2	11	8	-1
AUS	513	2	13	519	2	11	5	-2	DEU	493	3	13	499	3	16	5	3
JPN	510	3	14	514	3	14	3	0	AUS	492	2	14	501	2	12	9	-2
DEU	509	3	15	512	3	16	3	1	NLD	492	3	15	500	3	15	8	0
NLD	508	3	16	514	3	15	5	-1	BEL	491	3	16	496	2	17	5	1
BEL	508	2	17	510	2	17	2	0	ISR	487	4	17	503	4	9	16	-8
SVN	507	2	18	509	2	18	2	0	SVN	486	2	18	492	2	18	5	0
FRA	502	3	19	507	3	19	5	0	FRA	483	3	19	490	3	19	7	0
CZE	502	3	20	504	3	20	3	0	PRT	483	3	20	489	3	21	6	1
PRT	501	3	21	502	3	21	2	0	CZE	481	3	21	489	2	20	8	-1
CHE	496	3	22	499	3	23	3	1	AUT	479	3	22	485	3	22	6	0
AUT	495	3	23	500	3	22	4	-1	CHE	472	3	23	480	3	23	9	0
ITA	489	2	24	494	2	24	5	0	LVA	469	2	24	474	2	26	5	2
LVA	488	2	25	491	2	26	2	1	LTU	469	2	25	474	2	25	6	0
ESP	488	2	26	490	2	27	3	1	HUN	467	3	26	473	3	28	6	2
ISL	486	3	27	493	3	25	7	-2	ESP	467	2	27	472	2	29	5	2
HUN	485	2	28	488	2	28	3	0	ISL	466	3	28	477	3	24	11	-4
LTU	483	2	29	485	2	30	2	1	ITA	463	3	29	474	3	27	11	-2
LUX	482	2	30	486	2	29	4	-1	LUX	458	2	30	466	2	30	8	0
SVK	474	2	31	478	2	31	4	0	TUR	458	2	31	460	2	31	1	0
TUR	473	2	32	474	2	33	1	1	SVK	450	3	32	458	3	32	8	0
GRC	470	4	33	474	4	32	5	-1	GRC	445	4	33	456	3	33	11	0
CHL	464	3	34	465	3	34	2	0	CHL	441	3	34	446	3	34	5	0
MEX	429	3	35	430	3	35	0	0	MEX	412	3	35	413	3	35	1	0
COL	420	3	36	422	3	36	2	0	COL	405	3	36	408	3	36	3	0
Moy.	498	0,46		502	0,44				Moy.	479	0,48		486	0,45			

Note. Les moyennes surlignées en bleu sont supérieures à la moyenne arithmétique de l'OCDE ; celles en orange sont inférieures.

3.4. Effort et différences filles-garçons

Les différences de performances moyennes en fonction du genre, les ratios de variance et les tailles de l'effet sont présentées dans le tableau 7. Les différences de moyennes reflètent clairement un avantage en lecture des filles (statistiquement significatif dans tous les pays sauf en Colombie lors de la session 1), en moyenne de 25 points en session 1 et de 31 points en session 2. Après exclusion des élèves peu engagés dans le test, les écarts de performance se réduisent un peu, de près de 3 et 5 points respectivement. Les élèves désengagés étant plus fréquemment des garçons, leur faiblesse en lecture comparativement aux filles est surestimée.

Pareillement, la plus grande variabilité des performances en lecture des garçons s'explique en partie par leurs moindres efforts dans le test. En moyenne, la variance des garçons est 1,14 fois plus grande que celle des filles en session 1 et 1,19 fois celle des filles en session 2 ; au niveau pays, le ratio est statistiquement supérieur à 1 dans 20 et 32 pays respectivement. Après filtrage, les ratios moyens passent respectivement à 1,10 et 1,15 et les ratios deviennent non significatifs dans 11 et 10 pays respectivement. Cette diminution des écarts de variance s'explique également par les moindres efforts des garçons, qui entrainent une plus forte surestimation à gauche de leur distribution de compétence (élèves peu performants).

Finalement, les analyses relatives aux tailles de l'effet amènent des conclusions similaires à celles des différences de moyennes. Après filtrage, les écarts entre filles et garçons se réduisent, les tailles de l'effet moyennes passant de -0,27 à -0,25 (session 1) et de -0,32 à -0,28 (session 2). Les différences de genre en matière d'effort conduisent donc, en lecture, à une surestimation des différences de moyenne et de variabilité entre filles et garçons.

Tableau 7. Par session, différences de performances selon le sexe (différence de moyenne et taille de l'effet) et ratio de variance, avant et après filtrage. Pour chaque indice, par session, la moyenne de l'OCDE (μ) est présentée avec le nombre de pays présentant soit une différence statistiquement significative (+ ou -) ou non significative (NS) sur les différences/tailles de l'effet ; soit un ratio statistiquement supérieur ou inférieur à 1 (>1 ou <1) ou non significatif (NS)

	Session 1					**Session 2**				
Différence de moyennes	μ	-	- NS	+ NS	+	μ	-	-NS	+ NS	+
Brute	-25,53	35	1			-31,10	36			
Filtrée	-22,73	35	1			-26,31	36			
Ratio de variance	μ	<1	<1 NS	>1 NS	>1	μ	<1	<1 NS	>1 NS	>1
Brut	1,14		1	15	20	1,19			4	32
Filtré	1,10			27	9	1,15			14	22
Taille de l'effet	μ	-	- NS	+ NS	+	μ	-	-NS	+ NS	+
Brute	-0,27	35	1			-0,32	36			
Filtrée	-0,25	35	1			-0,28	36			

4. Conclusions

L'objectif de cette étude est d'identifier le manque d'effort dans le test de lecture de PISA 2018 et d'estimer son impact sur deux résultats phares de ces enquêtes. La technique du NT10 permet d'identifier les réponses rapides, celles-ci consistant plus souvent que les réponses typiques en des omissions ou des réponses erronées. Le pourcentage de réponses rapides tend à augmenter au fur et à mesure de la progression dans les étapes du test et est plus élevé pour les questions à réponse ouverte. Ces résultats, congruents avec la littérature, attestent de la validité de cette mesure pour refléter l'effort engagé dans le test. Par ailleurs, les élèves peu engagés, fournissant de moins bonnes réponses aux modules de routage, sont plus susceptibles d'être orientés vers des modules faciles à l'étape 2 ; la mesure de la compétence dans ces derniers modules est donc potentiellement plus entachée par le manque d'efforts que dans les modules difficiles. Lors de la troisième étape, aucun pattern quant à la relation entre effort et difficulté du module n'émerge, probablement parce que les réponses rapides peuvent y refléter un manque d'effort mais aussi de la fatigue ou un manque de temps.

L'effort a ensuite été estimé au niveau élève sur base des TR, et la corrélation de cette estimation avec les mesures d'effort auto-rapportées (effort investi) semble indiquer un lien faible, en moyenne de 0,19, dans la seconde session (la corrélation étant négligeable en session 1) : le thermomètre PISA semble donc mesurer surtout l'effort en deuxième et dernière session. De plus, toujours au départ des réponses au thermomètre PISA, il ressort que les élèves aux profils les moins motivés (irréalistes, démotivés et pragmatiques) ont en moyenne engagé le moins d'effort dans le test.

Le manque d'effort engendre une sous-estimation des performances moyennes nationales, d'autant plus que le désengagement est élevé. Les différences nationales d'effort influencent les classements des pays, en particulier lorsque les pays sont « dans un mouchoir de poche » ; cependant, la significativité des écarts à la moyenne OCDE reste stable après filtrage. Au sein des pays, les différences de performance entre groupes, telles celles en fonction du genre, peuvent

également être impactées par un différentiel d'effort. Ainsi, étant donné leurs moindres efforts, la faiblesse en lecture des garçons ainsi que leur plus grande variance sont surestimées.

Des implications pour le développement de tests et pour l'interprétation des résultats sont dégagées.

4.1. Implications pour le développement de tests à faibles enjeux

L'effort investi dans les épreuves à faibles enjeux peut influencer les résultats (et les conclusions) de ces enquêtes. Des dispositifs préventifs permettant de réduire le désengagement lors de la passation du test (Rios, 2021) existent et regroupent une diversité d'interventions : augmenter l'importance du test (en cotant ou en annonçant que le test sera noté [voir Fumel & Keskpaik, 2017] ou en informant les participants sur l'importance de leur participation et des résultats en découlant [voir Keskpaik & Rocher, 2012]), offrir des incitants (bons cadeaux ou récompenses financières par exemple, voir Braun et al., 2011), promettre aux étudiants un feedback sur leur performance, ou modifier le design de l'évaluation (modification des formats des items ou de leur difficulté par exemple). Ainsi, par exemple, Keskpaik et Rocher (2012) notent qu'en France, les établissements scolaires ont une faible connaissance du programme PISA, ce qui pourrait influer sur l'effort des élèves. Par ailleurs, pour les tests administrés sur ordinateur, un monitorage de l'engagement des élèves à travers la rapidité de leur réponse peut soit permettre au surveillant d'être informé des efforts des élèves en temps réel et, si besoin, d'intervenir (Wise et al., 2019), soit afficher un message d'alerte directement à l'étudiant. Enfin, la longueur du test influence l'effort investi par les élèves, ce dernier déclinant en fin de test ; une adaptation du nombre de questions posées et, en conséquence, du temps alloué pour le test, pourrait dès lors limiter le désengagement en fin de session.

En parallèle de ces mesures préventives, des mécanismes de détection du manque d'effort lors de l'analyse des données gagnent à être mis en œuvre. Ceux-ci passent par des mesures d'effort, basées sur les TR ou auto-rapportées : ces dernières quantifient cependant surtout l'effort investi juste avant leur administration, dans un effet de récence. Le moment d'administration de ces items auto-rapportés doit donc être réfléchi en amont, éventuellement en utilisant plusieurs temps de mesure (par exemple, à la fin de chaque session). De plus, le sérieux des réponses fournies au thermomètre gagne à être mis en lumière au regard des mesures d'effort basées sur les temps de réponse. Ainsi, les profils ayant réalisé le moins d'efforts selon l'indice RTE sont les élèves irréalistes, démotivés et pragmatiques en termes de réponses au thermomètre. Les élèves démotivés et pragmatiques semblent avoir déclaré dans le thermomètre un effort consistant avec leur effort tel que reflété par leurs temps de réponse ; à l'inverse, les élèves irréalistes (ayant déclaré avoir fourni plus d'effort à PISA qu'à un test comptant pour le bulletin) pourraient avoir répondu de manière également peu motivée au thermomètre PISA. La littérature propose plusieurs hypothèses quant à l'origine de ces réponses irréalistes au thermomètre : désirabilité sociale ou incompréhension des échelles dues à un faible niveau de lecture (Butler & Adams, 2007), ou encore expression de la frustration des élèves quant aux évaluations cotées quotidiennement pratiquées ou motivation liée au contexte non ordinaire de ces épreuves (Dierendonck et al., 2013). L'analyse des temps de réponses met en exergue une hypothèse supplémentaire, ces élèves irréalistes pouvant avoir été peu engagés dans le test cognitif mais aussi dans le thermomètre PISA, leur réponse irréaliste au thermomètre traduisant potentiellement un manque d'effort également lors de leur réponse à ce dernier.

Par ailleurs, dans un test adaptatif, le manque d'effort des élèves interfère avec le processus de routage. Les élèves désengagés sont davantage dirigés vers des modules plus faciles en raison de leurs faibles performances. Les biais engendrés par le désengagement peuvent être plus élevés dans les modules faciles ; ces derniers, en particulier ceux situés en milieu de test, méritent donc une attention accrue car ils concentrent plus de répondants peu motivés. Cependant, la moindre

difficulté des items subséquents pourrait éventuellement inciter les élèves fournissant peu d'efforts à s'engager dans le test. Alors que les *testings adaptatifs* visent à augmenter la précision de la mesure, en particulier aux extrémités de la distribution de la compétence (Zenisky et al., 2010), cet objectif semble plus atteint dans les modules relativement difficiles.

Enfin, d'après Silm et al. (2020), la relation entre l'effort et le score au test semble plus forte avec des élèves plus âgés, et l'effort engagé dans le test tendrait à diminuer avec l'âge (Keskpaik & Rocher, 2012 ; Weis et al., 2017). Le biais lié à l'effort est donc potentiellement moindre en primaire (par exemple, dans les données PIRLS ou du PASEC) qu'en secondaire (données PISA par exemple), voire dans l'enseignement supérieur ou auprès d'adultes.

4.2. Implications concernant les comparaisons de groupes de répondants

Un différentiel d'effort peut altérer la comparaison des performances entre groupes. Au sein d'une population, si un groupe est moins engagé, alors sa performance sera plus sous-estimée. Par exemple, les différences de performance en lecture selon le genre sont exacerbées par les moindres efforts des garçons. L'effet du désengagement sur l'estimation de l'écart est ainsi fonction de la différence d'effort mais aussi du sens de la différence de performance. Par exemple, en mathématiques (domaine traditionnellement mieux réussi par les garçons, voir par exemple Reilly et al., 2015), les moindres efforts des garçons engendreraient une sous-estimation de leur avantage scolaire dans ce domaine.

Au niveau international, des différences d'effort national peuvent altérer les comparaisons d'indicateurs d'efficacité et d'équité. Les écarts de moyenne mais aussi de dispersion peuvent être impactés. Le désengagement gonfle artificiellement la queue gauche des distributions, accroissant la variance estimée et, comme illustré pour les différences filles-garçons, les écarts de variance peuvent être exacerbés. Il est dès lors recommandé d'analyser et de tenir compte de l'effort investi dans les tests afin de maximiser l'utilisabilité des résultats de ces études.

4.3. Implications pour l'interprétation des résultats

L'analyse des différences de performance avant et après filtrage a dégagé des tendances quant à l'effet du désengagement sur les résultats. Concernant le classement des pays sur base de leur performance moyenne, les résultats après filtrage dévoilent un classement différent. Ce constat souligne la relative fragilité de ces palmarès internationaux au regard de l'erreur de mesure, fragilité par ailleurs déjà documentée notamment au regard de l'erreur d'échantillonnage, nombre de différences entre pays voisins dans le classement n'étant pas significatives (Champollion & Barthes, 2012 ; Duru-Bellat, 2019).

Si le classement international a été ébranlé par le filtrage sur base de l'effort, relevons que ce n'est pas le cas de la significativité des écarts à la moyenne OCDE. Ces résultats, qui par ailleurs tiennent comptent de l'erreur d'échantillonnage, permettent donc une interprétation et une utilisation beaucoup plus fiable que les classements. Les résultats de cette étude s'inscrivent donc dans la lignée des constats relatifs aux limites des classements internationaux : ils invitent à sortir d'une lecture en termes de palmarès et à adopter une approche approfondie des résultats. Cette étude ne discrédite pas les résultats des enquêtes PISA (les pays performants restant performants après filtrage, par exemple) : elle en souligne la robustesse tout en montrant la fragilité des interprétations qui se fondent sur les classements.

4.4. Limites et perspectives

Plusieurs limites de cette étude sont à mentionner. Tout d'abord, les mesures d'effort basées sur les TR reposent sur l'identification d'un seuil en deçà duquel une réponse est jugée rapide et désengagée. Elles détectent donc les réponses désengagées rapides mais pas les lentes (Wise & Kong, 2005) ; si le manque d'effort vient d'un manque d'intérêt pour le test, on peut cependant émettre l'hypothèse que les élèves souhaiteront terminer au plus vite le test plutôt que d'en éterniser la passation. Il existe de nombreuses méthodes pour fixer ce seuil (voir par exemple Rios & Deng, 2021). La méthode du NT10 offre l'avantage d'être facilement implémentable et requiert moins d'individus que, par exemple, les seuils basés sur des modèles de mixture ; d'autres méthodes pourraient néanmoins identifier des élèves potentiellement différents. Relevons que cet article utilise le TR moyen d'un item afin de calculer le seuil, comme recommandé dans la littérature ; les distributions des TR étant asymétriques, le TR médian pourrait aboutir à des seuils différents et de futures recherches pourraient analyser l'effet du choix de la mesure de tendance centrale sur l'effort estimé. De plus, les seuils sont calculés séparément pour chaque session ; la comparabilité entre sessions est donc limitée. Si une mesure d'effort devait être calculée à travers sessions, un seuil commun à celles-ci devrait être utilisé ; un tel seuil serait probablement plus strict pour la première session et moins sévère pour la seconde session que ceux employés dans cet article.

De plus, les mesures basées sur les TR peuvent également confondre désengagement et manque de temps[4]. La diminution de l'effort en fin de session peut confondre le désengagement avec un manque de temps. Ainsi, l'écart entre les modules faciles et difficiles en termes de pourcentages de désengagement observé dans l'étape 2 ne se reproduit pas au module suivant, les pourcentages de réponses rapides étant élevés aussi dans les modules difficiles. Ceci peut traduire un désengagement chez les élèves performants répondant aux modules difficiles (ces élèves recevant un test calibré pour être plus difficile que si le test avait été linéaire), mais aussi un manque de temps en fin de test.

Par ailleurs, la relativement faible ampleur des corrélations entre les mesures d'effort basées sur les TR et auto-rapportées pourrait s'expliquer par des biais de réponse dans ces dernières. Ainsi, un élève qui aurait fait de son mieux pour répondre aux questions mais qui sait qu'il a échoué pourrait dire qu'il a engagé peu d'efforts pour attribuer son échec à de la désinvolture ; inversement, des élèves peu engagés pourraient volontairement surestimer leur effort par effet de désirabilité sociale (Finn, 2015 ; Wise & Kong, 2005).

En outre, le seuil (25%) utilisé pour le filtrage des élèves désengagés est relativement indulgent, la littérature utilisant généralement des seuils plus stricts (10%) ; ce seuil plus clément permet de ne supprimer que les élèves fortement désengagés. Dans tous les cas et indépendamment de la valeur utilisée pour le filtrage, ce dernier repose sur l'hypothèse qu'il n'y a pas de lien entre effort et compétence. Si ce n'était pas le cas, alors le filtrage des élèves désengagés biaiserait l'échantillon et surestimerait, par exemple, la performance moyenne au lieu de produire des statistiques corrigées pour le manque d'effort. Les estimations filtrées reposent donc sur l'hypothèse d'une indépendance entre effort et compétence ; à l'inverse, les estimations brutes, non filtrées, pourraient être valides dans un contexte de corrélation parfaite entre effort et compétence. De plus, la motivation à engager des efforts dans le test pourrait ne pas être liée directement à la compétence des élèves mais plutôt découler de leur motivation et intérêt, de manière générale, pour l'école. Considérant qu'une réponse à un item pour laquelle l'élève n'a pas engagé d'effort constitue une absence de mesure (Wise, 2017) et donc une donnée manquante, les trois

[4] Les mesures d'effort basées sur les TR sont basées sur les travaux de Schnipke et Scrams (1997) qui, dans un test à forts enjeux (pour lequel les élèves sont, théoriquement, engagés dans le test), analysent le manque de temps à travers la production de réponses rapides.

configurations précitées peuvent être rattachées à trois mécanismes distincts : les données manquantes de façon aléatoire (MAR, *missing at random*), de façon complètement aléatoire (MCAR, *missing completely at random*) et de façon non aléatoire (MNAR, *missing not at random*). Ces trois cas de figure supposent donc des relations distinctes entre la compétence, les déterminants de la compétence et le fait de répondre de manière non engagée (assimilé à une donnée manquante) :

- MCAR : l'engagement dans le test n'est lié ni à la compétence, ni à ses déterminants. La probabilité d'observer une réponse manquante (non engagée) à un item est totalement indépendante des autres variables ainsi que de la réponse qui aurait été fournie à l'item si l'élève y avait répondu (Enders, 2010).
- MNAR : la compétence de l'élève détermine son engagement dans le test. Un élève n'engage pas d'effort car il sait qu'il sera de toute façon dans l'incapacité de répondre (la relation entre la probabilité de survenue d'une donnée manquante à un item, soit un manque d'effort, et la compétence serait donc négative).
- MAR : un déterminant de la compétence (l'intérêt pour l'école ou le genre par exemple) est également déterminant de la motivation au test. Ainsi, un élève peu intéressé par l'école pourrait ne pas vouloir engager d'effort dans la passation du test PISA (et aurait donc une plus grande probabilité d'exhiber des comportements désinvestis en comparaison de collègues plus intéressés par l'école).

Plus de recherches sont cependant requises pour clarifier le lien entre compétence, déterminants de la compétence et effort. Enfin, si des techniques de filtrage ou de modélisation des réponses désengagées sont employées, une réflexion doit être menée quant à la communication aux répondants de l'emploi de ces techniques. Les élèves font souvent l'hypothèse que toute réponse non correcte est cotée 0 et, s'ils venaient à être informés que les réponses désengagées ne sont pas prises en compte dans l'estimation de leur performance, ils pourraient modifier leur stratégie de réponse au test : ils pourraient attribuer encore moins de valeur à leur performance, choisir les items auxquels ils souhaitent répondre ou encore arrêter de répondre lorsqu'ils estiment avoir « engrangé » assez de bonnes réponses.

5. Références bibliographiques

American Educational Research Association, American Psychological Association b, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing (ED565876)*. https://eric.ed.gov/?id=ED565876

Baye, A., & Monseur, C. (2016). Gender differences in variability and extreme scores in an international context. *Large-Scale Assessments in Education, 4, Article 1*. https://doi.org/10.1186/s40536-015-0015-x

Braun, H., Kirsch, I., & Yamamoto, K. (2011). An experimental study of the effects of monetary incentives on performance on the 12th-grade NAEP reading assessment. *Teacher college record, 113*(11), 2309-2344. https://doi.org/10.1177/016146811111301101

Butler, J., & Adams, R. J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of applied measurement, 8*(3), 279-304.

Cattonar, B., & Mangez, E. (2014). Codages et recodages de la réalité scolaire. *Revue internationale d'éducation de Sèvres, 66*, 61-70. https://doi.org/10.4000/ries.3999

Champollion, P., & Barthes, A. (2012). De l'usage et du mésusage du classement par rang en matière de médiatisation de l'évaluation internationale PISA. *Questions Vives, 6*. https://doi.org/10.4000/questionsvives.895

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & practice in assessment, 8,* 69-82.

Dierendonck, C., Milmeister, M., Milmeister, P., Weis, C., Fischbach, A., Ugen, S., & Martin, R. (2016). L'implication des élèves lors d'évaluations externes à faibles enjeux : le cas de l'évaluation « Épreuves Standardisées » au Luxembourg. In C. Cavaco, N. Alves, P. Guimaraes, C. Dierendonck, P. Alves, A. Machado, P. Rodrigues, M. Marques & C. Paulos, *Actes du 28e colloque de l'ADMEE-Europe : évaluations et apprentissages* (pp. 546-548). ADMEE-Europe.

Dierendonck, C., Milmeister, M., Weis, C., & Milmeister, P. (2017). Motivation et effort des élèves lors des évaluations externes à faibles enjeux : une question de validité et de mesure. In *Actes du 29e colloque de l'ADMEE-Europe : l'évaluation, levier pour l'enseignement et la formation. Réseau thématique : apprentissages scolaires et évaluations externes* (pp. 3-17). ADMEE-Europe.

Dierendonck, C., Sonnleitner, P., Ugen, S., Keller, U., Fischbach, A., & Martin, R. (2013). *La mesure de la motivation et de l'effort des élèves dans le cadre des Épreuves Standardisées au Luxembourg*. Congrès de l'actualité de la recherche en éducation et formation (AREF-AECSE), Montpellier, France.

Duru-Bellat, M. (2019). Évaluations, mesures ou classements ? À propos des enquêtes PISA. *Revue française de linguistique appliquée, 24*(1), 7-19. https://doi.org/10.3917/rfla.241.0007

Eklöf, H. (2010). Skill and will: test-taking motivation and assessment quality. *Assessment in Education Principles Policy Practice, 17*(4), 345–356. https://doi.org/10.1080/0969594X.2010.516569

Enders, C. K. (2010). *Applied missing data analysis*. The Guilford Press.

Finn, B. (2015). Measuring motivation in low-stakes assessments (research report RR-15-19). *Educational Testing Service*.

Fumel, S., & Keskpaik, S. (2017). La motivation des élèves à répondre à un test standardisé : résultats d'une étude dans le cadre de Cedre compétences langagières et littératie. *Éducation & formations, 93*, 105-119.

Grey, S., & Morris, P. (2018). PISA: multiple 'truths' and mediatised global governance. *Comparative Education, 54*(2), 109-131. https://doi.org/10.1080/03050068.2018.1425243

Keskpaik S., & Rocher, T., (2012). Les évaluations à faibles enjeux : quel rôle joue la motivation ? Une expérience à partir de PISA. In *24e colloque de l'ADMEE-Europe, L'évaluation des compétences en milieu scolaire et en milieu professionnel. Résumés des ateliers (Ateliers 1-14)* (pp. 100-107). ADMEE-Europe.

Keskpaik, S., & Rocher, T. (2015). La motivation des élèves français face à des évaluations à faibles enjeux. Comment la mesurer ? Son impact sur les réponses. *Éducation & formations, 86-87*, 119-139.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement, 67*(4), 606–619. https://doi.org/10.1177/0013164406294779

Kunter, M., Schumer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., et al. (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente*. Max-Planck-Institiut für Bildungsforschung.

Lindner, M. A., Lüdtke, O., & Nagy, G. (2019). The onset of rapid-guessing behavior over the course of testing time: A matter of motivation and cognitive resources. *Frontiers in Psychology, 10, Article* 1533. https://doi.org/10.3389/fpsyg.2019.01533

Mead, A. D. (2006). An introduction to multistage testing. *Applied Measurement in Education, 19*(3), 185-187. https://doi.org/10.1207/s15324818ame1903_1

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 international results in reading*. TIMSS & PIRLS International Study Center. http://timssandpirls.bc.edu/pirls2016/international-results/

OCDE. (2009). *PISA data analysis manual: SAS (2nd ed.)*. OECD publishing. https://doi.org/10.1787/9789264056251-en

OCDE. (n.d.) *PISA 2018 technical report*. https://www.oecd.org/pisa/data/pisa2018technicalreport/

OCDE. (2015). *The ABC of gender equality in education: Aptitude, behaviour, confidence*. OECD Publishing. https://doi.org/10.1787/9789264229945-en

OCDE. (2019a). *PISA 2018 Assessment and Analytical Framework*. OECD Publishing. https://doi.org/10.1787/b25efab8-en

OCDE. (2019b). *Résultats du PISA 2018 (Volume I) : Savoirs et savoir-faire des élèves*. Éditions OCDE. https://doi.org/10.1787/ec30bc50-fr

OCDE. (2019c). *Pisa 2018 results (volume II): Where all students can succeed*. OECD Publishing. https://doi.org/10.1787/b5fd1b8f-en

Petersen, J. (2018). Gender difference in verbal performance: A meta-analysis of United States state performance assessments. *Educational Psychology Review, 30*, 1269-1281. https://doi.org/10.1007/s10648-018-9450-x

Pons, X. (2010). Qu'apprend-on vraiment de Pisa ? Sociologie de la réception d'une enquête internationale dans trois pays européens (2001-2008). *Revue internationale d'éducation de Sèvres, 54*, 51-59. https://doi.org/10.4000/ries.850

Reilly, D., Neumann, D. L., & Andrews, G. (2015). Sex differences in mathematics and science achievement: A meta-analysis of national assessment of educational progress assessments. *Journal of Educational Psychology, 107*(3), 645-662. http://dx.doi.org/10.1037/edu0000012

Reilly, D., Neumann, D. L., & Andrews, G. (2019). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). *American Psychologist, 74*(4), 445-458. http://dx.doi.org/10.1037/amp0000356

Rios, J. (2021). Improving test-taking effort in low-stakes group-based educational testing : a meta-analysis of interventions. *Applied Measurement in Education, 34*(2), 85-106. https://doi.org/10.1080/08957347.2021.1890741

Rios, J. A., & Deng, J. (2021). Does the choice of response time threshold procedure substantially affect inferences concerning the identification and exclusion of rapid guessing responses? A meta-analysis. *Large-Scale Assessments in Education, 9, Article 18*. https://doi.org/10.1186/s40536-021-00110-8

Schnipke, D. L., & Scrams, D. J. (1997). Modeling item response time with a two-state mixture model: A new method of measuring speededness. *Journal of Educational Measurement, 34*(3), 213-232. http://www.jstor.org/stable/1435443

Silm, G., Pedaste, M., & Täht, K. (2020). The relationship between performance and test-taking effort when measured with self-report or time-based instruments: A meta-analytic review. *Educational Research Review, 31*, Article 100335. https://doi.org/10.1016/j.edurev.2020.100335

Weis, C., Milmeister, P., Milmeister, M., & Dierendonck, C. (2017) Dans quelle mesure les élèves font-ils les tests ÉpStan sérieusement ? In S. Ugen, R. Martin, & A. Fischbach, *Les épreuves standardisées : comment sont-elles perçues par les acteurs concernés ? Principaux constats et clarifications* (pp. 19-26). Université du Luxembourg.

Wigfield, A., & Eccles, J. A. (1992). The development of achievement task values: a theoretical analysis. *Developmental Review, 12*(3), 265-310. https://doi.org/10.1016/0273-2297(92)90011-P

Wigfield, A., & Eccles, J. A. (2000). Expectancy-value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68–81. https://doi.org/10.1006/ceps.1999.1015

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education, 19*(2), 95–114. https://doi.org/10.1207/s15324818ame1902_2

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation and implications. *Educational Measurement: Issues and Practice, 36*(4), 52-61. https://doi.org/10.1111/emip.12165

Wise, S. L. (2020) Six insights regarding test-taking disengagement. *Educational Research and Evaluation, 26*(5-6), 328-338. https://doi.org/10.1080/13803611.2021.1963942

Wise, S. L., & Gao, L. (2017). A general approach to measuring test-taking effort on computer-based tests. *Applied Measurement in Education, 30*(4), 343-354. https://doi.org/10.1080/08957347.2017.1353992

Wise, S. L., & Kong, X. (2005). Response time effort: a new measure of examinee motivation in computer-based tests. *Applied Measurement in Education, 18*(2), 163-183. https://doi.org/10.1207/s15324818ame1802_2

Wise, S. L., Kuhfeld, M. R., & Soland, J. (2019). The effects of effort monitoring with proctor notification on test-taking engagement, test performance, and validity. *Applied Measurement in Education, 32*(2), 183-192. https://doi.org/10.1080/08957347.2019.1577248

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: implications for test development and measurement practice. *Applied Measurement in Education, 22*(2), 185-205. https://doi.org/10.1080/08957340902754650

Zenisky, A., Hambleton, R. K., & Luecht, R. M. (2009). Multistage testing: issues, designs, and research. Dans van der Linden, W., & Glas, C. (Eds), *Elements of adaptive testing. Statistics for social and behavioral sciences* (pp 355–372). Springer. https://doi.org/10.1007/978-0-387-85461-8_18