Psychological Test and Assessment Modeling, 2024.66:50-62
DOI: https://doi.org/10.2440/001-0012

The impact of filtering out rapid-guessing examinees on PISA 2015 country rankings

Michalis P. Michaelides[1], *Militsa G. Ivanova*[1], *Demetris Avraam*[2]

[1] Department of Psychology, University of Cyprus
[2] Department of Public Health, Policy and Systems, University of Liverpool

Abstract:
International large-scale assessments are low-stakes tests for examinees and their motivation to perform at their best may not be high. Thus, these programs are criticized as invalid for accurately depicting individual and aggregate achievement levels. In this paper, we examine whether filtering out examinees who rapid-guess impacts country score averages and rankings. Building on an earlier analysis that identified rapid guessers using two different methods, we re-estimated country average scores and rankings in three subject tests of PISA 2015 (Science, Mathematics, Reading) after filtering out rapid-guessing examinees. Results suggest that country mean scores increase for all countries after filtering, but in most conditions the change in rankings is minimal, if any. A few exceptions with considerable changes in rankings were observed in the Science and Reading tests with methods that were more liberal in identifying rapid guessing. Lack of engagement and effort is a validity concern for individual scores, but has a minor impact on aggregate scores and country rankings.

Keywords: Rapid guessing, response time effort, PISA, filtering

Correspondence:

Michalis P. Michaelides

Institutional address: Dept. of Psychology, 1 Panepistimiou Avenue, 2109 Aglantzia, P.O. Box 20537, 1678 Nicosia, Cyprus

Email: Michaelides.michalis@ucy.ac.cy

Introduction

Scores on achievement tests are considered valid indicators of individual achievement, assuming that the test-takers were sufficiently motivated to take the test and actively engaged with the test content (Eklöf, 2010). Examinees' test-taking effort in achievement tests has been found to have significant impact both on their performance and on the psychometric properties of the test (Wise & DeMars, 2005). In low-stakes international large-scale assessments (ILSA), where test-takers face minimal or no personal consequences, they may not be motivated to perform at their best (Lee & Chen, 2011). For instance, examinees who respond to a test item rapidly without paying attention to the item content are expected to perform worse on average, compared to those who engage in a solution behavior (Wise, 2017).

Examinee behavior at the item level can be classified as rapid guessing if a response is given very fast, without the test-taker reading and thinking about the item question; responses given after a certain time threshold are considered as solution behavior (Wise & Kong, 2005). Much of the literature has been concerned with different ways of distinguishing rapid guessing from solution behavior via a threshold, as summarized in Wise (2017); a threshold could be fixed for all items comprising a test, or item-specific determined visually or through modeling, based on item characteristics, or normative sample data.

Wise and Kong (2005) proposed an approximate indicator of overall effort on an entire test, termed response time effort (RTE), as the proportion of item-level responses for which an examinee exhibited solution behavior. In a metanalysis by Silm, et al. (2020) RTE was found to be highly associated to test performance, with a much higher effect size than self-reports of effort.

Studies with ILSA data have confirmed this strong correlation between RTE and achievement scores on tests (Michaelides et al., 2020; Michaelides & Ivanova, 2022; Pools & Monseur, 2021). Examinees who engage in more rapid guessing behavior perform lower than their peers, thus, their individual scores may underestimate their achievement levels and invalidate test outcomes. Consequently, aggregate scores, such as those at the country level, may be biased downwards and if there are cultural differences in test-taking effort, the validity of country-level comparisons may be harmed (Debeer et al., 2014; Goldhammer et al., 2016).

At the same time, ILSA results may have significant implications for jurisdictions and institutions, making them high-stakes at the policy level. When confronted with unfavorable performance outcomes and declines in scores, politicians, teacher unions, and other interested stakeholders seek to explain the low performance. One plausible justification for suboptimal country performance is the low interest and effort exhibited by test-takers. However, initial empirical evidence about test-taking effort does not support this claim, at least not universally. In interviews with Norwegian PISA participants stated that they were generally motivated to do their best in the low-stakes PISA study (Hopfenbeck & Kjærnsli, 2016). Test motivation and performance on a

short version of the PISA test did not significantly change even after experimentally manipulating the stakes of testing in German 9th-graders (Baumert & Demmrich, 2001). In self-reports of intended effort and test behaviors across vignettes of varying personal stakes, Zhao et al. (2020, 2022) found no differences with Shanghai students, and some differences in New Zealand students. Finally, Gneezy et al. (2017) showed that experimental manipulation of incentives in achievement tests, had an effect on scores in high schoolers in the US, but not in Shanghai.

In the context of cross-country comparisons, studies have documented variations in test-taking effort across countries when examinees self-report the effort they invest on a test in surveys of Trends in International Mathematics and Science Study (TIMSS; Eklöf, et al., 2014) and Programme for International Student Assessment (PISA; Eklöf, 2015). When item response times, which are considered less susceptible to response biases, are employed as measures of effort, engagement, or rapid guessing, country differences have also been reported in the Programme for the International Assessment of Adult Competencies (Goldhammer, et al., 2016) and PISA (Azzolini, et al., 2019; Guo & Ercikan, 2020; Michaelides & Ivanova, 2022). Many of these studies have additionally highlighted heterogeneous associations between the measure of effort used and test performance across country samples.

A separate question is whether differential test-taking effort by examinees across countries has an influence on aggregate scores (Zamarro et al., 2019) and on country rankings. Rios and Guo (2020) found that despite differential noneffortful responding in four countries which were administered a college-level critical thinking test, filtering out noneffortful responses did not change country rankings. Guo and Ercikan (2020) studied nine PISA participating jurisdictions and found no substantial impact on country rankings when examinees with low response time effort indicators were excluded.

The purpose of this paper was to examine the impact of filtering out examinees with low test-taking effort, operationalized as low RTE, on country average scores and rankings in a large and comprehensive sample of PISA jurisdictions; countries participating in the 2015 computerized administration of PISA Science, Mathematics, and Reading assessments were included. Response time data by item were used to identify rapid guessers in two ways: a fixed 5-second and a more liberal 15% normative threshold. Based on the strong association of RTE with test performance, we hypothesized that country averages would increase after filtering out rapid guessers, especially with the more liberal threshold. Although the average increases would be different for each country after the filtering, we did not anticipate noticeable changes in country rankings since all would shift upwards.

Method

Sample

The study relies on a sample of 56 PISA 2015 countries or jurisdictions that administered the computerized version of the assessment and was analyzed in Michaelides and Ivanova (2022)[1]. A two-stage sampling design was implemented within nearly all countries: at least 150 schools with 15-year-old students were sampled and then 42 students were typically selected from each school (OECD, 2017). While a minimum sample of 5250 students was targeted in each country, the country samples ranged from 3371 to 23141 for Science which was the major subject in PISA 2015. Not all students responded to subject tests that were not major in the assessment cycle. Country-specific sample sizes ranged from 1396 to 9288 for Mathematics and from 1374 to 9317 for Reading.

Measures

The PISA data used for the study are freely available at https://www.oecd.org/pisa/data/2015database/. The computerized administration of PISA 2015 included a two-hour assessment: one hour was devoted to Science, and one hour on one or two different subjects, Mathematics, Reading, or Collaborative Problem Solving (OECD, 2017a), hence the sample sizes for each subject were different. The "cognitive items total time/visits data file" provides the total time spent by a student on each item. As described in Michaelides and Ivanova (2022), the variables for total time spent on an item were used to obtain two measures of rapid-guessing behavior at the item level: (a) a fixed 5-second threshold to flag very rapid responses (including rapid omits) and (b) an item-specific normative threshold based on the 15 % (NT15) of the item mean response time (Wise, 2019) in the country sample. For each examinee, the overall test-taking effort in the subject test was calculated as the proportion of responses on which a student did not engage in rapid-guessing (Wise & Kong, 2005). Therefore, each examinee had two alternative RTE scores based on the fixed and the NT15 thresholds. We considered the fixed 5-second threshold, an indicator also utilized in PISA reports, as an identical cut-off for all items which ignores item features and provides a conservative way of flagging rapid guessers (i.e., increased risk for Type II error). The normative threshold is item specific and depends on item and sample features, which is pertinent for PISA items that tend to be complex, occasionally long and of varying types. The NT15 is also more liberal and allows

[1] Three jurisdictions, Massachusetts, North Carolina, and Spain (Regions), were excluded from the sample of 59 jurisdictions analyzed by Michaelides & Ivanova (2022), so the sample in the current paper is 56. Following a reviewer's suggestion we extended the analysis in the current paper to include the Science test, which was not studied in the 2022 paper.

for identifying more rapid guessers, considering both the probability of identification of false positive and false negative lack of effort (Linder et al., 2017).

Under a matrix sampling design, each examinee responds to a subset of the total item pool (OECD, 2017b). PISA estimates ten plausible values for each individual in every subject test using item response theory scaling as a proficiency estimate. The "cognitive item data file" includes scored variables for each item response and the ten plausible values per domain for each examinee, which were used as achievement measures.

Statistical Analysis

For each country, we merged the individual RTE scores with the corresponding plausible values. The final student sampling weight variable provided in the PISA student-level dataset was used in all subsequent analysis to allow for representative estimates of student proficiency. The average score and the rank for each country was estimated.

In the filtering analysis, we removed students who had (a) an RTE < 1, which means that they responded rapidly on at least one item, and (b) an RTE < .95, which means that they responded rapidly on more than 5 % of the items in their test. RTE was estimated in two ways: fixed 5-seconds, and NT15. The analysis was performed separately for the Science, Mathematics, and Reading tests. After each filtering method, the average scores and ranks for all countries were re-estimated.

Finally, because Mathematics and Reading were not the major subjects in the assessment, not all students within a country received tests in Mathematics and in Reading. However, PISA reports plausible values for all students via imputation. We repeated our filtering analysis using the complete country samples, and separately using only those students who actually received the Mathematics or the Reading test. Data files and R analysis code in R version 4.3.1 (R Core Team, 2023) are available in OSF https://osf.io/m2pna/.

Results

The percentage of examinees identified for rapid guessing differed by country. The average percentages and their standard deviations (SD) by threshold method and criterion can be seen on Table 1. As expected, more rapid guessers were identified with the liberal NT15 than with the fixed 5-second threshold and – trivially – with an RTE < 1 criterion than with an RTE < .95. With Science being the longer test taken by all examinees in a country, larger percentages of rapid guessers were found compared to the other two subjects under all conditions. Under the fixed threshold, 6.84 % and 12.73 % examinees per country were on average flagged as rapid guessers; under the NT15, the estimates more than doubled, showing more extensive rapid guessing behavior. In the other two subjects, the percentages of rapid guessers were small under the fixed 5-second threshold and about twice as high for Reading than for Mathematics. The percentages were similar for the two subject tests, with the NT15 threshold, but much larger than the stricter 5-second threshold. Finally, when only the examinees who took each test were considered (excluding those with imputed plausible values) rapid guessing was higher ranging from 3.39 % to 24.61 % depending on the method used.

Table 1.
Mean and standard deviation of the percentage of examinees filtered out from each country sample under different thresholds and criteria

	Based on the entire country sample				Based on examinees who took the test (excluding imputed scores)			
	NT15<.95	NT15<1	Fixed<.95	Fixed<1	NT15<.95	NT15<1	Fixed<.95	Fixed<1
Science assessment								
Mean	17.16	27.05	6.84	12.73	n/a			
SD	6.57	7.82	4.46	6.15				
Mathematics assessment								
Mean	6.44	10.43	1.46	2.61	15.16	24.61	3.39	6.05
SD	2.21	2.71	1.17	1.79	4.53	5.49	2.40	3.64
Reading assessment								
Mean	6.78	10.27	3.20	4.94	15.94	24.20	7.43	11.52
SD	2.92	3.42	2.10	2.77	6.15	7.13	4.30	5.65

Note. NT15 = 15% Normative Threshold, SD = Standard Deviation

When examinees identified as rapid guessers were filtered out and country mean scores were re-estimated, all country scores increased. Average increases were larger in Science, followed by Reading, and smaller for Mathematics, where less rapid guessing was detected; larger increases were found under the NT15 than the fixed 5-second threshold (Table 2). The average changes were very small in most cases, considering that the PISA scale has a standard deviation of 100 points. Standard deviations of the increase were also small, suggesting minor differences in the impact of the filtering process on country averages. A noteworthy average increase that could exceed a fifth of a standard deviation was observed under the liberal NT15 threshold in Science, as well as in the other two subjects when examinees with imputed scores were excluded from the analysis.

Table 2.
Average increase (and standard deviation) in the country mean score after filtering out rapid guessing examinees under different thresholds and criteria

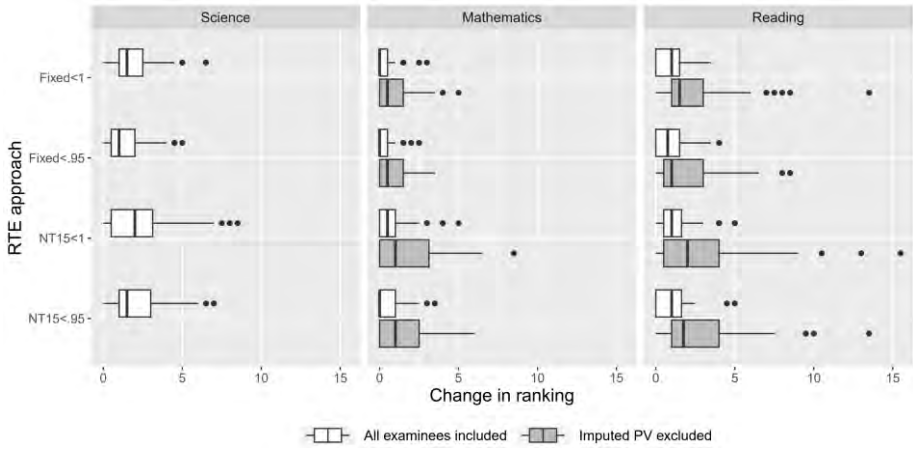
	Based on the entire country sample				Based on examinees who took the test (excluding imputed scores)			
	NT15<.95	NT15<1	Fixed<.95	Fixed<1	NT15<.95	NT15<1	Fixed<.95	Fixed<1
Science assessment								
Mean	17.77	25.22	7.35	12.14	n/a			
SD	7.60	9.20	5.14	6.89				
Mathematics assessment								
Mean	5.43	7.05	1.44	2.28	14.31	20.09	3.48	5.63
SD	2.19	2.49	1.14	1.55	5.76	7.31	2.61	3.69
Reading assessment								
Mean	6.71	8.28	3.48	4.70	17.94	23.72	8.68	12.14
SD	2.85	2.94	2.32	2.72	7.81	8.97	5.57	6.95

Note. NT15 = 15% Normative Threshold, SD = Standard Deviation

To further investigate whether these score increases are consequential, we re-estimated the ranks of countries after filtering out rapid guessing examinees. Generally, in the list of 56 countries, changes (in absolute value) were small, especially under the fixed 5-second threshold compared to the NT15, which flags more rapid guessers (Figure 1). In Science, there were small changes in rankings after filtering; the median change was between one and two depending on the criterion. The modal change in ranking was one (or two under the NT15 < 1 criterion). The largest changes were observed in France, Israel, and Sweden (increase) and Ireland (decrease). In Mathematics, when the entire country samples were considered (white boxplots in Figure 1) changes were minimal with a modal change of zero and a median of less than one

under all thresholds and criteria. When examinees with imputed plausible values were excluded (grey boxplots in Figure 1), the change in rankings after filtering was slightly higher with a mode of zero and a median change of at most one. The largest change was observed in the French sample (increase). In Reading, changes were more noticeable: in the entire samples the modal change was typically zero and the median change was at most one. When the examinees with imputed plausible values were excluded the modal change was at most one, and the median change at most two; however, for a few countries, there were larger changes in Reading rankings: Israel and France (increase) and Macao (decrease).

Figure 1.
Boxplots for the change in country rankings by subject, threshold and criterion



Note. White boxes show change in rankings when the entire country sample was considered and grey boxes when only examinees who took the test (excluding imputed plausible values) was considered. RTE = Response Time Effort.

Discussion

Several technical characteristics in the design of ILSA guarantee reliable and valid scores: large, random samples to ensure representativeness within countries, matrix sampling to allow for broad content coverage for each subject, multiple item formats to evaluate many content areas, proficiency levels and skills, adaptive testing in digital administration for more efficient measurement. Reduced test-taking effort and disengagement with the assessments constitute however, a considerable threat to score validity (Debeer et al., 2014; Wise & Demars, 2010). Unmotivated examinees

underperform on the test, and therefore aggregate scores do not reflect country performance accurately. While reasonable, this claim has not been empirically tested. Some recent studies with a small number of countries suggested that filtering out rapid-guessing examinees did not impact country rankings (Rios & Guo, 2020; Guo & Ercikan, 2020). The current study included 56 countries that participated in the computerized administration of PISA 2015 for a more comprehensive test of the claim.

Not surprisingly, removing examinees who provide rapid responses on some test items, leads to increased country average scores. This holds for all countries, since rapid guessing can be found globally in low-stakes ILSA (Michaelides & Ivanova, 2022). There are cross-country variations in the rates of rapid guessing, which may be reflected in differential increases in country mean scores, but our findings showed that these differences were not large. Hence, the impact of filtering out rapid guessers is low in terms of rankings. Unlike the two studies which found no impact of filtering (Rios & Guo, 2020; Guo & Ercikan, 2021), we documented some changes, minor in almost all cases, because we considered a much larger number of countries in our comparison. Some countries have average scores that are very similar, so differential increases due to filtering, however small, may result in minor changes in rankings. Minimal effects of disengagement on aggregate statistics on a state summative assessment in the US were also reported by Wise et al. (2021).

The literature on test-taking effort and engagement is large with multiple suggestions on how to quantify this construct. After certain operationalizations based on self-reports, behavioral indices based on item response time became more common as a result of the growth of computerized programs. There are however limitations in whether item response time can adequately describe effort, due to misclassifications that may occur: an examinee who responds quickly as a "rapid guesser" may be very skilled test-wise and knowledgeable, while one who does not respond rapidly is not really investing effort into attentively constructing a response. We limited our investigation to rapid guessing, defined as a rapid recording of a response below a certain threshold. We considered two types of thresholds: a strict, fixed 5-second one that minimizes misclassifications of rapid guessers, and a more liberal, item-specific normative threshold (15% of mean item response time). As expected, the NT15 resulted in slightly larger changes than the fixed one. This implies that the way examinees are labeled as rapid guessers and filtered out can be consequential in the outcomes of the assessment.

Another positive aspect of the analysis was the filtering criterion: excluding any examinee who rapid guessed even just one item, or more than 5% of the items. This sensitivity check revealed minor differences as well, with increased impact when methodological decisions flag more examinees as rapid guessers. Listwise deletion of examinees based on indications of item-level rapid guessing, leads to inflated means when ability correlates with careless responding (Rios et al., 2017); moreover, in the case of studies with sampling procedures designed to ensure representativeness to a population, it may impair the inferences drawn from the purified samples.

Instead of filtering out rapid guessing, test-taking behavior can alternatively be conceptualized as an informative variable relevant to real-life behavior. With data from just three countries, Pohl et al. (2021) have shown that if speed and response propensity are included in the estimation of test outcomes along with ability estimates will result in changes in country rankings. This alternative idea also highlights the dependency of results on how a construct like test-taking behavior is operationalized, as well as how behaviors like speed and response propensity are weighted together with ability in producing a test outcome.

The study of the impact of rapid guessing filtering on rankings would benefit by replicating this analysis with additional studies, or datasets, e.g., TIMSS. Further research on alternative ways of classifying examinees as disengaged, like clickstreams or number of actions (Ivanova et al., 2020; Ivanova & Michaelides, 2023; Tang et al., 2023) would also help generalize the robustness of this finding. Finally, the minor impact of filtering out rapid guessing examinees on country rankings shown in this paper, should not be interpreted as inconsequential. Test-taking effort and engagement with the test are more consequential at the individual examinee level (Wise et al., 2021). Ways to reduce rapid guessing and increase attentive engagement with the test (cf. Lee & Chen, 2011) would dramatically benefit individual achievement, in low- as well as in high-stakes testing situations.

Declarations

Availability of data and materials. Original data are available on the PISA 2015 website. Analysis code files and auxiliary data files are available at OSF: https://osf.io/m2pna/?view_only=36ef1f60b36548f98abe92ecec1a90a6

Competing interests. The authors declare no competing interests.

Funding. Funding for this work was provided by the A.G. Leventis Foundation to MPM.

Acknowledgments. The authors would like to thank Andreas Economides for assistance with programming.

Authors' contributions. MPM and MGI jointly devised the project. MPM wrote the Introduction and Methods, designed and conducted the analysis and drafted the Results, and Discussion. MGI reviewed and edited the entire manuscript. DA wrote the code and prepared the datasets for the analysis, and reviewed the manuscript.

List of abbreviations

ILSA: International Large-Scale Assessments

NT15: Normative threshold 15 %

OECD: Organization for Economic Co-operation and Development

PISA: Programme for International Student Assessment

RTE: Response Time Effort

TIMSS: Trends in International Mathematics and Science Study

References

Azzolini, D., Bazoli, N., Lievore, I., Schizzerotto, A., & Vergolini, L. (2019). *Beyond achievement. A comparative look into 15-year-olds' school engagement, effort, and perseverance in the European Union.* European Commission. https://doi.org/10.2766/98129

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*(3), 441-462. https://doi.org/10.1007/BF03173192

Debeer, D., Buchholz, J., Hartig, J., & Janssen, R. (2014). Student, school, and country differences in sustained test-taking effort in the 2009 PISA reading assessment. *Journal of Educational and Behavioral Statistics, 39*(6), 502-523. https://doi.org/10.3102/1076998614558485

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

Eklöf, H. (2015). Swedish students' reported motivation and effort in PISA, over time and in comparison with other countries. In *To respond or not to respond: The motivation of Swedish students in taking the PISA test* (pp. 11-60). Swedish National Agency for Education.

Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2014). A cross-national comparison of reported effort and mathematics performance in TIMSS advanced. *Applied Measurement in* Education, *27*(1), 31-45. https://doi.org/10.1080/08957347.2013.853070

Gneezy, U., List, J. A., Livingston, J. A., Qin, X., Sadoff, S., & Xu, Y. (2019). Measuring success in education: the role of effort on the test itself. *American Economic Review: Insights, 1*(3), 291-308.

Goldhammer, F., Martens, T., Christoph, G, & Lüdtke, O. (2016). *Test-taking engagement in PIAAC. OECD Education Working Papers*, No. 133, OECD Publishing. http://dx.doi.org/10.1787/5jlzfl6fhxs2-en

Guo, H., & Ercikan, K. (2020). Differential rapid responding across language and cultural groups. *Educational Research and Evaluation, 26*(5-6), 302-327. https://doi.org/10.1080/13803611.2021.1963941

Hopfenbeck, T. N., & Kjærnsli, M. (2016). Students' test motivation in PISA: The case of Norway. *The Curriculum Journal*, *27*(3), 406-422. https://doi.org/10.1080/09585176.2016.1156004

Ivanova, M. G., & Michaelides, M. P. (2023). Measuring test-taking effort on constructed-response items with item response time and number of actions. *Practical Assessment, Research, & Evaluation, 28*(15). Available online: https://doi.org/10.7275/pare.1921

Ivanova, M. G., Michaelides, M. P., & Eklöf, H. (2020). How Does the Number of Actions on Constructed-Response Items Relate to Test-Taking Effort and Performance? *Educational Research and Evaluation, 26*(5-6), 252-274. https://doi.org/10.1080/13803611.2021.1963939

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling, 53*(3), 359-379.

Lindner, M. A., Lüdtke, O., Grund, S., & Köller, O. (2017). The merits of representational pictures in educational assessment: Evidence for cognitive and motivational effects in a time-on-task analysis. *Contemporary Educational Psychology, 51*, 482-492. https://doi.org/10.1016/j.cedpsych.2017.09.009

Michaelides, M.P., & Ivanova, M. (2022). Response time as an indicator of test-taking effort in PISA: country and item-type differences. *Psychological Test and Assessment Modeling, 64*(3), 304-338.

Michaelides, M. P., Ivanova, M., & Nicolaou, C. (2020). The Relationship between Response-Time Effort and Accuracy in PISA Science Multiple Choice Items. *International Journal of Testing, 20*(3), 187-205. https://doi.org/10.1080/15305058.2019.1706529

OECD. (2017a). *PISA 2015 Assessment and Analytical Framework: Science, Reading, Mathematic, Financial Literacy and Collaborative Problem Solving*. OECD Publishing. https://doi.org/10.1787/9789264281820-en.

OECD. (2017b). *PISA 2015 technical report*. OECD Publishing. https://www.oecd.org/pisa/data/2015-technical-report/

Pohl, S., Ulitzsch, E., & von Davier, M. (2021). *Reframing rankings in educational assessments. Science, 372(*6540), 338-340. https://www.science.org/doi/10.1126/science.abd3300

Pools, E., & Monseur, C. (2021). Student test-taking effort in low-stakes assessments: Evidence from the English version of the PISA 2015 science test. *Large-scale Assessments in Education, 9*(1), 1-31. https://doi.org/10.1186/s40536-021-00104-6

R Core Team (2023). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rios, J. A., & Guo, H. (2020). Can culture be a salient predictor of test-taking engagement? An analysis of differential noneffortful responding on an international college-level assessment of critical thinking. *Applied Measurement in Education, 33*(4), 263-279. https://doi.org/10.1080/08957347.2020.1789141

Rios, J. A., Guo, H., Mao, L., & Liu, O. L. (2017). Evaluating the Impact of Careless Responding on Aggregated-Scores: To Filter Unmotivated Examinees or Not? *International Journal of Testing, 17*(1), 74–104. https://doi.org/10.1080/15305058.2016.1231193

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*, 100335. https://doi.org/10.1016/j.edurev.2020.100335

Tang, S., Samuel, S., & Li, Z. (2023). Detecting Atypical Test-Taking Behavior with Behavior Prediction Using LSTM. *Psychological Test and Assessment Modeling, 65*(1), 76-124.

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. (2019). An information-based approach to identifying rapid-guessing thresholds. *Applied Measurement in Education, 32*(4), 325-336. https://doi.org/10.1080/08957347.2019.1660350

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1-17. https://doi.org/10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2010). Examinee noneffort and the validity of program assessment results. *Educational Assessment, 15*(1), 27-41. https://doi.org/10.1080/10627191003673216

Wise, S. L., Im, S., & Lee, J. (2021). The impact of disengaged test taking on a state's accountability test results. *Educational Assessment, 26*(3), 163-174. https://doi.org/10.1080/10627197.2021.1956897

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

Zamarro, G., Hitt, C., & Mendez, I. (2019). When students don't care: Reexamining international differences in achievement and student effort. *Journal of Human Capital, 13*(4), 519-552. https://doi.org/10.1086/705799

Zhao, A., Brown, G. T., & Meissel, K. (2020). Manipulating the consequences of tests: How Shanghai teens react to different consequences. *Educational Research and Evaluation, 26*(5-6), 221-251. https://doi.org/10.1080/13803611.2021.1963938

Zhao, A., Brown, G. T., & Meissel, K. (2022). New Zealand students' test-taking motivation: An experimental study examining the effects of stakes. *Assessment in Education: Principles, Policy & Practice, 29*(4), 397-421. https://doi.org/10.1080/0969594X.2022.2101043