Psychological Test and Assessment Modeling, Volume 64, 2022 (3), 304-338

# Response time as an indicator of test-taking effort in PISA: country and item-type differences

*Michalis P. Michaelides*[1] *& Militsa Ivanova*

Department of Psychology, University of Cyprus

**Abstract**
In low-stakes assessments, when test-takers do not invest adequate effort, test scores underestimate the individual's true ability, and ignoring the impact of test-taking effort may harm the validity of test outcomes. The study examined the level of examinees' test-taking effort and accuracy in the Programme for International Student Assessment (PISA) across countries and different item types. The 2015 PISA computerized assessment was administered in 59 jurisdictions. Behavioral measures of students' test-taking effort were constructed for the Mathematics and Reading assessments by applying a fixed and a normative threshold on item response times to identify rapid guessing. The proportion of rapid guessers on each item was found to be small on average, about 3 %, according to the normative and 1 % with a fixed five-second threshold. Rapid guessing was about twice as high in human-coded open-response items, compared to simple and complex multiple-choice items, and computer-scored open-response items. Average performance for rapid guessers was on average much lower than for test-takers engaged in solution behavior for all types of items and more pronounced in Reading than in Mathematics. Weighted response time effort indicators by country were very high, and positively correlated with country mean PISA score. No other robust correlates were found with response time effort at the country level. Computerized test administrations facilitate the use of response time as a proxy for examinee test-taking effort. Programs may monitor this behavior to identify cross-country differences prior to comparisons of performance and for developing interventions to promote engagement with the assessment.

**Keywords:** Test-taking behavior, rapid guessing, response time effort, PISA.

[1] *Correspondence concerning this article should be addressed to* Michalis P. Michaelides, Dept. of Psychology, 1 Panepistimiou Avenue, 2109 Aglantzia, P.O. Box 20537, 1678 Nicosia, Cyprus, E-mail: Michaelides.michalis@ucy.ac.cy

## Introduction

Test-takers participate in testing events and obtain scores that reflect their performance level. To perform well, sufficient knowledge and skills are required, but they also need to be motivated enough to actively engage with the test content (Eklöf, 2010); otherwise, their lack of effort will be reflected in the score. Across studies, test-taking effort and performance correlate moderately to strongly (Silm, Pedaste, & Täht, 2020). In a metanalytic comparison motivated examinees' test performance was .58 standard deviations higher than that of unmotivated ones (Wise & DeMars, 2005). Low test-taking effort had also a significant impact on the psychometric properties of achievement tests: reducing test validity (Wise, 2009) and inflating reliability coefficients (Wise & DeMars, 2009). The problem of the biased test results due to low test-taking effort, may be loom larger in low-stakes assessment programs, such as the Programme for International Student Assessment (PISA), where few or no consequences are posed to test-takers for low performance (Rutkowski & Wild, 2015). Hence, there has been an increased research interest in studying test-taking effort in low-stakes assessments.

## Measuring Test-taking Effort

Verbalization (e.g., self-reports, interviews) and behavioral measures (e.g., response time, choice of task, response missingness, change in performance) have been suggested as indicators for test-taking effort (Eklöf, 2010). For instance, a number of PISA administrations have used the effort thermometer, a two-item post-test self-report instrument assessing the effort students have invested on the test (Organization for Economic Co-operation and Development [OECD], 2015, 2019). Self-report effort measures are simple, economical, quick for administration, flexible, and easy to incorporate in tests and surveys; however, they are susceptible to response bias, rely on examinees' motivation, honesty and ability to understand the questions, and tend to be global, not item-level, measures of individuals' intentions or perceptions of invested effort on the whole test (Eklöf, 2010; Wise, 2015).

Computerized tests facilitate the collection of response times and other types of process data, revealing aspects of examinees' response behavior during assessments. Response time has been proposed as a valid behavioral indicator of test-taking effort (Wise & Kong, 2005). Examinees willing to invest effort when responding to an item will engage in solution behavior by spending time looking for the correct response, while disengaged test-takers will present rapid-guessing or rapid-omit behavior by giving a fast response or rapidly skip the item without actually considering it (Wise & Kong, 2005; Wise & Gao, 2017). Since response time provides information about examinee test-taking behavior at the item level, it allows researchers to track possible changes in effort during the test session (Wise & Kingsbury, 2016). In addition to the item level, response time can be used to denote effortful behavior on the whole test by calculating the proportion of items in an assessment for which an examinee exhibited solution behavior, an indicator known as Response Time Effort (RTE; Setzer, Wise, van den Heuvel, & Ling, 2013). Response times are less vulnerable than verbalizations to response bias, cultural differences in response style, or examinee ability to understand a self-report questionnaire. Such a conceptualization of effort captures rapid responding, but not all disengaged behavior. However, effort approximated with response time indicators is strongly associated with test performance (Michaelides, Ivanova, & Nicolaou, 2020; Pools & Monseur, 2021), while self-reported effort is not (Eklöf & Knekta, 2017); in a metanalysis by Silm et al. (2020) the average correlations were .72 and .33, respectively.

## Threshold Identification when Using Response Time for Effort Indicators

The use of response time as a measure of effort requires a pre-defined threshold to identify disengaged test-taking behavior. Examinees responding in a time faster than the threshold are considered as rapid guessers, while the rest of the responses are recorded as solution behaviors (Wise & DeMars, 2006). In the theory of rapid guessing behavior, Wise (2017) stipulated that the proportion of correct responses on an item, item accuracy, will be higher for solution-based examinees than for rapid guessers.

Various methods have been proposed to establish a threshold, including using: a common threshold for all items (Wise, Ma, Kingsbury, & Hauser, 2010), a normative threshold (i.e., a percentage of the average time spent by examinees on a particular item; Wise & Ma, 2012), a cumulative proportion method combining the response accuracy rate with response time (Guo et al., 2016), a two-class finite mixture model (Schnipke & Scrams, 1997), or item response theory modeling (Ulitzsch, Penk, von Davier, & Pohl, 2021; Ulitzsch, von Davier, & Pohl, 2020). A threshold can be identified also based on: item length (Wise & Kong, 2005), visual inspection of time frequency distribution (Setzer et al., 2013), or examination of the change in item-total correlation (i.e., the correlation between item accuracy and test performance; Wise, 2019).

Previous literature has come to ambiguous results regarding the preferred threshold identification method (Kong, Wise, & Bhola, 2007; Wise, 2019), as there are strengths and weaknesses for each one (cf., Soland, Kuhfeld, & Rios, 2021). Kong et al. (2007) recommended using the information available about the test items as a criterion to select a threshold estimation method. Thresholds based on items length or visual inspection of response time distribution may not be suitable for achievement tests using large and frequently changing item pools. The OECD (2019) has used a common threshold of 5 seconds across items, along with a self-report effort scale (i.e., the effort thermometer) to describe students' test-taking effort in PISA 2015 and PISA 2018 administrations. However, an item specific threshold identification method, such as the Normative Threshold (NT), may be more suitable than the use of a common threshold for all items in a test containing a combination of different type of items from various difficulty levels, and produce more viable thresholds than the visual method, finite mixture models, or the cumulative proportion method (Soland et al., 2021). Approaches based on item response theory are promising because model-based estimates at the item and respondent level are used, but they rest on restrictive assumptions (Ulitzsch, et al., 2021; Ulitzsch, Pohl, Khorramdel, Kroehne, & von Davier, 2022) and encounter convergence problems under various sample and item conditions (Ulitzsch, et al., 2020).

Each threshold identification method leads to different levels of misclassification error. Trying to reduce the possibility of false-positive results increases the possibility of false-negatives (Wise, 2017). More conservative threshold identification methods (Kroehne, Deribo, & Goldhammer, 2020), such as 10 % Normative Threshold (NT10) may be a useful criterion for proctor notification or invalidation of test results due to low effort, while more liberal thresholds, such as NT15 or NT20, have been recommended in research interested mainly in aggregated test scores (Soland, et al., 2021; Wise, 2019; Wise & Kuhfeld, 2021). The use of NT15 has been preferred by Lindner, Lüdtke, and Nagy (2019), since it has been shown to balance the possibility of false-positive and false negative errors in the threshold identification.

## The Relationship of Item Characteristics and Context Factors with Test-taking Effort

Item position has been frequently studied as a critical characteristic affecting examinee effort; overall, effort tends to decrease towards the end of a testing session (Debeer, Buchholz, Hartig, & Janssen, 2014; Goldhammer, Martens, & Lüdtke, 2017; Pools & Monseur, 2021). Item characteristics such as less amount of reading material, more answer options, and inclusion of graphics in the item were associated with less rapid-guessing behavior (Setzer, Wise, Heuven, & Ling, 2013; Wise, Pastor, & Kong, 2009). DeMars (2000) presented evidence for higher item non-response and lower motivation in low-stakes constructed response, compared to MC items. However, selected response items elicited primarily rapid guessing and occasional rapid omits, while for short answer and constructed response items disengaged participants either rapidly omitted items or quickly entered “rapid perfunctory” responses (Wise & Gao, 2017). Additional empirical evidence regarding effort on different item types, including novel formats administered in computerized assessments is needed, especially in low-stakes settings.

In terms of test type, examinee effort was found to be lower for longer and more difficult tests (Barry & Finney, 2016), Reading than Mathematics tests (Wise et al., 2010), Problem Solving and Literacy than Numeracy tests (Goldhammer, Martens, Christoph, & Lüdtke, 2016), and for tests carrying low or no consequences for the test-takers (Wise, Kingsbury, Thomason, & Kong, 2004). Effort was not significantly influenced by the time of the year, or by the day of the week a test is administered, unlike the time of the day the test is taken, with the solution behavior occurring earlier in the day and diminishing as the day progresses, regardless of the student grade or the subject examined (Mathematics or Reading; Wise et al., 2010).

## Cross-national Comparisons

A common concern in discussions for international large-scale assessments with performance rankings, is that low level of effort invested by examinees may invalidate country-level comparisons (Debeer et al., 2014; Goldhammer et al., 2016). Swedish 12th-graders in TIMSS 2008 Advanced demonstrated lower self-reported test-taking effort, poorer test performance, and a stronger relationship between test-taking effort and performance compared to Norwegian and Slovenian samples; no significant differences across the three countries were observed when comparing only students reporting high level of test-taking effort (Eklöf, Pavešič, & Grønmo, 2014). Rapid response behavior was different across four nations in an international college-level assessment of critical thinking, but without changing the relative ranking of countries after filtering (Rios & Guo, 2020). Cross-national differences in effort may also be dependent on the subject examined. Larger cross-country discrepancies in test-taking engagement, measured via item response times, was found in problem solving than in Literacy and Numeracy in the Programme for the International Assessment of Adult Competencies (PIAAC; Goldhammer, et al., 2016). However, cross-cultural findings should be cautiously interpreted because studies have reported cultural differences in response tendencies in self-report scales (van de Vijver & He, 2014) and in response time scales (Shin, Kerzabi, Joo, Robin, & Yamamoto, 2020).

Cross-national variations in test-taking effort have been observed in PISA as well. Countries reporting a very high level of effort on the “effort thermometer”, had low average performance in PISA 2012, while countries with slightly lower (but still high) level of self-reported effort included some of the best performers on the test (i.e., Taipei). The relationship between self-reported test-taking effort and performance in PISA 2012 varied across countries: some nations exhibited positive relationships with different magnitude between the two variables, while others presented a lack of relationship or even negative associations between examinee effort and test results (Eklöf, 2015). In PIAAC, a positive association between effortful behavior as measured with response times and average country performance was reported by Goldhammer et al. (2016). When comparing European Union countries on PISA 2015, Azzolini, Bazoli, Lievore, Schizzerotto, and Vergolini (2019) found considerable country-level heterogeneity; Hungarian students presented the lowest, while Portuguese 15-year olds invested the highest level of test-taking engagement. Finally, among nine different language and cultural groups in PISA 2018 Science, rapid response rates and their association with performance were different (e.g., the association was stronger in the Arab countries and weaker in the Chinese sample), but with limited impact on country rankings (Guo & Ercikan, 2021). In general, there is lack of evidence regarding country-level characteristics and their relationship with operationalizations of test-taking effort, as well as descriptive differences on RTE across a large number of jurisdictions participating in PISA. To the extent that there are differences between jurisdictions in RTE, then there is ground for concern about the validity of inferences for score comparability: if there are substantial differences in the level of effort across countries, then country average scores will be differentially impacted. Gneezy et al. (2017) suggested that country scores in low-stakes assessments reflect differences both in ability and in test-taking effort. They presented experimental evidence that manipulation of incentives in achievement testing, has an effect on test scores in some cultures, but not in others.

## Purpose of the Study

The purpose of the study was to describe test-taking effort in the Mathematics and Reading PISA 2015 assessments, as measured by examinees' item response time, and its relationship with test performance at the country level. Proportions of effortless test-taking behavior were compared for 59 jurisdictions which administered the computerized PISA assessment. By using response time, a behavioral indicator of test-taking effort, the study aims to provide evidence on whether there are differences in the extent of rapid guessing (and rapid omitting) across countries, and thus inform the validity of cross-country comparisons. A fixed (5 seconds) and a 15 % item-specific normative threshold (NT15) were considered in determining effortless responses. Moreover, the levels of rapid guessing behavior were compared across types of items administered in PISA. We hypothesized that students engaging in effortful behavior will be more accurate than students identified as rapid guessers, irrespective of item type. Country-level RTE was expected to differ across countries and to positively correlate with country performance. Finally, we explore whether country-level of effort is associated with other aggregate characteristics: (a) average country score in mathematics, may be related to average effort (Debeer et al., 2014; Eklöf et al., 2014), along with (b) geographic region, (c) Gross Domestic Product, and (d) country-level personality scores; individual students' effort was found to relate with conscientiousness, agreeableness and openness to experience (Barry, Horst, Finney, Brown, & Kopp, 2010; Barry & Finney, 2016), but at the country level such associations have not been studied.

## Method

### Sample

PISA is an international assessment program administered triennially, which aims to evaluate the level of preparation of 15-year-old students to meet the challenges of modern society. Over 500 000 15-year-olds from 35 OECD countries and 37 partner countries and economies participated in PISA 2015. The assessment was administered on computer in 56 countries and one jurisdiction (i.e., B-S-J-G, China), while the remaining economies administered the paper-and-pencil test (OECD, 2017a). In the cognitive item data file (available at https://www.oecd.org/pisa/data/2015database/), three additional jurisdictions were included (Massachusetts, North Carolina, and Spain Regions). The final sample for analysis in the current study consisted of 59[1] countries or jurisdictions.

For each country, a two-stage stratified sample design[2] was used. In the first stage, at least 150 schools were selected in each country from a school sampling frame, containing all schools with 15-year-old students. The probability of selecting a school to participate was proportional to school size. In the second stage, from all eligible students in a selected school, 42 were typically sampled in countries administering a computer-based assessment (OECD, 2017a). Generally, a minimum sample size of 5250 students were selected to participate in the computer-based assessment in each country (OECD, 2017a). A subsample of examinees participated in the Mathematics and Reading assessments, since these subjects were not the major focus of PISA 2015. The number of examinees by jurisdiction and assessment can be seen on Tables A1 for Mathematics and A2 for Reading in the Appendix.

## Measures

PISA 2015 examined students' knowledge and skills in the areas of Science, with Reading, Mathematics and Collaborative Problem-solving being minor areas of assessment; Financial Literacy was an optional domain. The assessment consisted of two one-hour sessions with a 5-minute break in between. Each test session included two 30-minute clusters of test material. Students got two clusters assessing their knowledge in the major subject, and two clusters focusing on one or two of the minor subjects assessed. The PISA 2015 items pool consisted of six new Science clusters, six clusters from each of Science, Reading, and Mathematics to measure trends, and three Collaborative Problem-solving clusters. Following test completion, the students were administered a 35-minute background questionnaire with information about themselves, their home, and their school (OECD, 2017b).

A total of 810 minutes of testing material was included in the PISA 2015 survey and a different combination of items was given to different students under a matrix sampling design.

The test was a combination of MC questions and constructed-response items, which were arranged in groups called units based on different text scenario or graphic

[1] Data for Malaysia were stored in a separate file due to concerns during data adjudication and were not analyzed. Data for Cyprus were also stored separately and were merged with the international databased.

[2] A three-stage sample design was used in the Russian Federation; the additional stage comprised selection of geographical areas.

describing real-life situations (OECD, 2017b). Students could move back and forth between items within a unit but not between units (OECD, 2017a). The MC questions were divided into simple and complex MC based on the response format required by the item. Simple MC items usually required a selection of a single correct response from four response options or choosing an element from a graphic or text. A total of 20 Mathematics and 36 Reading simple MC items were included in PISA 2015 item pool. Complex MC items required responses to a series of "Yes/No" questions, selection of multiple responses from a list, completion of sentences with multiple gaps from a "drag-and-drop" menu, and relocation of elements to complete a matching or ordering task. A total of 14 Mathematics and 12 Reading complex MC items were included in the PISA 2015 item pool. Open-response items (OR) required written response or a graph drawing. The PISA 2015 OR items pool consisted of 47 Mathematics (26 computer-scored and 21 human-scored), and 52 Reading (7 computer-scored and 45 human-coded) items.

Scored variables and process data (i.e., response time, number of actions made by students while interacting with an item) were included in the cognitive datasets for each item separately. Variables ending with T represented the time students have spent during their last visit of an item, while variables ending with TT, which were included in a separate datafile,[3] reflected the total time participants have spent on a particular item (OECD, 2017a). Assuming that at least some amount time is required to meaningfully engage with an item, total item response latency was used to determine whether a student provided a response as a result of effort, or rapidly guessed and moved on to the next item. Item accuracy was decided from the scored variables; full credit was recorded as correct response.

Multiple types of missing variables have been identified in the cognitive PISA dataset: not presented to student, omitted by the student, invalid responses (e.g., responses that did not fit to the response format), not applicable (response provided to an item that should have been skipped), and valid skips (skipped item that should have been skipped). Omitted responses at the end of any of the one-hour sessions were coded as not reached. While omitted responses in the beginning or in the middle of a test session have been treated as wrong, not reached items have been treated as not administered by PISA (OECD, 2017a). For the current study, omitted and invalid responses were considered treated as wrong; all other types were retained as missing.

The following country-level variables were also included as correlates of country level RTE: Country average PISA Math/Reading score (OECD, 2016), the 2015 Gross Domestic Product and 2014 Human Development Index variables, Hofstede's six dimensions model of national culture (power distance, individualism, masculinity, uncertainty avoidance, long-term orientation, indulgence)[4], as well as country aggregates

[3] More information can be found in "Annex K: Uses and Reporting of Process Data" at https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018-TechReport-Annex-K.pdf

[4] From the dimension data matrix in http://www.geerthofstede.nl/

for the Big Five (McCrae & Costa, 1987) model[5] (extraversion, conscientiousness, openness to experience, agreeableness, neuroticism).

### Estimation of Thresholds and Response Time Effort

Classifying a response to an item as rapid guessing or solution behavior required a threshold on the item response time variable. Two types of thresholds have been used in the study: (a) For a fixed threshold of 5 seconds, item responses given within 5 seconds were classified as rapid guesses, while the remaining responses were considered solution behaviour; and (b) a liberal 15 % Normative Threshold (NT15) was selected over other item-specific threshold identification methods because of its potential to provide a valid threshold for each item (Wise, 2019). When using the NT15, the mean response time of an item was calculated for each country separately; then, 15 % of the mean response time was set as a threshold to distinguish rapid guessing from solution behavior.

The overall test-taking effort expended by a student on a particular subject was approximated with the student-level response time effort (RTE) score, which is defined as the proportion of responses on which a student engaged in solution behavior (Wise & Kong, 2005). RTE scores were calculated for each country and subject separately after applying sampling weights, the “final trimmed nonresponse adjusted student weights”. Data processing and visualizations has been conducted with tidyverse (version 1.3.2) and PerformanceAnalytics packages in R (version 4.2.0). The codes are available at: https://osf.io/pxq6h/

## Results

### Findings from the Mathematics Assessment

Across all countries and items, the number of rapid guessing test-takers under the fixed 5-second threshold was small, on average 4.84 students (*SD* = 6.66) per item. With the NT15 method, the average threshold across all items and all jurisdictions was 15.44 seconds (*SD* = 1.53) with more rapid guessers identified: a mean of 26.71 per item (*SD* = 23.92). In terms of percentages the extent of rapid guessing was small, 0.58 % and 3.26 % (averaged across items) for the two threshold methods (Table 1). The country with the highest percentages was Qatar with 2.97 % and 10.17 % respectively. Percentages for all countries are presented in the Appendix Table A1. Rates of

[5] Average Big Five scores for each country were obtained from the Big Five Personality Inventory (includes 56 countries) (Schmitt, Allik, McCrae, & Benet-Martínez, 2007) and the Revised NEO Personality Inventory (includes 51 countries) (McCrae, Terracciano, & Personality Profiles of Cultures Project, 2005).

rapid guessing were also estimated for different item types. A similar percentage of test-takers were classified as rapid guessers with simple and complex MC, and OR computer-scored items, about 0.50 % (fixed) and 2.50 % (NT15). The corresponding percentages were higher for OR human-coded items, 0.89 % and 5.39 % (Table 1).

Examinees who respond rapidly below a threshold on an item are less likely to perform well compared to those who spend more time engaging with the item. Averaging across all items and jurisdictions, this was clearly evident since the average percent correct for rapid guessers was low 0.07 (fixed) and 0.09 (NT15) compared to 0.42 and 0.43, respectively for solution-behavior students (Table 2). Similar differences were found when items types were analyzed separately.  Simple MC items had overall the highest accuracy indices; the average percent correct was more than 0.50 for solution behavior examinees and close to 0.20 for rapid guessers and both thresholds. Performance was lower for other types of items, with the lowest on OR human-coded items. However, on all types of items average percent correct for rapid guessers was very low overall and substantially lower than the corresponding accuracy indices for solution behavior examinees.

**Table 1.**
Extent of rapid guessing behavior across types of items under two threshold methods for Mathematics and Reading

| | Mathematics | | Reading | |
|---|---|---|---|---|
| | Fixed threshold | Normative threshold | Fixed threshold | Normative threshold |
| Descriptive statistics | Average (SD) | Average (SD) | Average (SD) | Average (SD) |
| Number of rapid guessers on all items | 4.84 (6.66) | 26.71 (23.92) | 9.98 (11.19) | 24.65 (20.41) |
| Percentage of rapid guessers | | | | |
| Across all items | 0.58 (0.47) | 3.26 (1.40) | 1.27 (0.87) | 3.13 (1.50) |
| On Simple MC | 0.53 (0.61) | 2.32 (1.41) | 1.56 (1.23) | 2.77 (1.57) |
| On Complex MC | 0.51 (0.39) | 2.63 (1.44) | 0.69 (0.54) | 2.07 (1.18) |
| On Open Response CS | 0.40 (0.34) | 2.60 (1.19) | 0.76 (0.57) | 2.71 (1.33) |
| On Open Response HC | 0.89 (0.59) | 5.39 (1.87) | 1.27 (0.83) | 3.78 (1.79) |

*Note*. MC=multiple-choice, CS=computer-scored, HC=human-coded, SD=standard deviation.

**Table 2.**
Average percent correct indices by type of items and test-taking behavior for the two threshold methods for Mathematics and Reading.

| | Mathematics | | | | | | | | Reading | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Fixed threshold | | | | Normative threshold | | | | Fixed threshold | | | | Normative threshold | | | |
| | Rapid guessers | | Solution behavior | | Rapid guessers | | Solution behavior | | Rapid guessers | | Solution behavior | | Rapid guessers | | Solution behavior | |
| Item group | Mean | (SD) | Mean | (SD) | Mean | (SD) | Mean | (SD) | Mean | (SD) | Mean | (SD) | Mean | (SD) | Mean | (SD) |
| All items | 0.07 | (0.04) | 0.42 | (0.09) | 0.09 | (0.02) | 0.43 | (0.09) | 0.08 | (0.04) | 0.56 | (0.07) | 0.11 | (0.03) | 0.57 | (0.07) |
| Simple MC | 0.19 | (0.09) | 0.55 | (0.09) | 0.22 | (0.06) | 0.56 | (0.09) | 0.18 | (0.06) | 0.66 | (0.06) | 0.22 | (0.05) | 0.66 | (0.06) |
| Complex MC | 0.05 | (0.06) | 0.45 | (0.11) | 0.12 | (0.03) | 0.46 | (0.11) | 0.06 | (0.06) | 0.39 | (0.08) | 0.12 | (0.05) | 0.39 | (0.08) |
| OR-CS | 0.03 | (0.05) | 0.41 | (0.09) | 0.04 | (0.04) | 0.42 | (0.09) | 0.03 | (0.06) | 0.45 | (0.06) | 0.04 | (0.05) | 0.46 | (0.06) |
| OR-HC | 0.02 | (0.03) | 0.29 | (0.09) | 0.02 | (0.02) | 0.30 | (0.10) | 0.01 | (0.02) | 0.55 | (0.08) | 0.02 | (0.03) | 0.56 | (0.08) |

*Note*. MC=multiple-choice, CS=computer-scored, OR=open-response, HC=human-coded, SD=standard deviation.

In Figure 1, the average percent correct for each item type can be seen for each country separately. The pairs of boxplots show that in all cases, rapid guessers (left boxplot on each pair) performed on average lower than solution behavior test-takers (right boxplot on each pair). This is consistent within all jurisdictions as can be seen by the grey lines that connect the indices for each one. Distributions for percent correct centered near zero for rapid guessers in OR items under both threshold methods; they were also less variable under the NT15 method for all types of items. As regards solution behavior examinees, accuracy was higher than for rapid guessers, but substantially variable across jurisdictions. Also, proportion correct indices tended to be higher in MC than in OR formats.

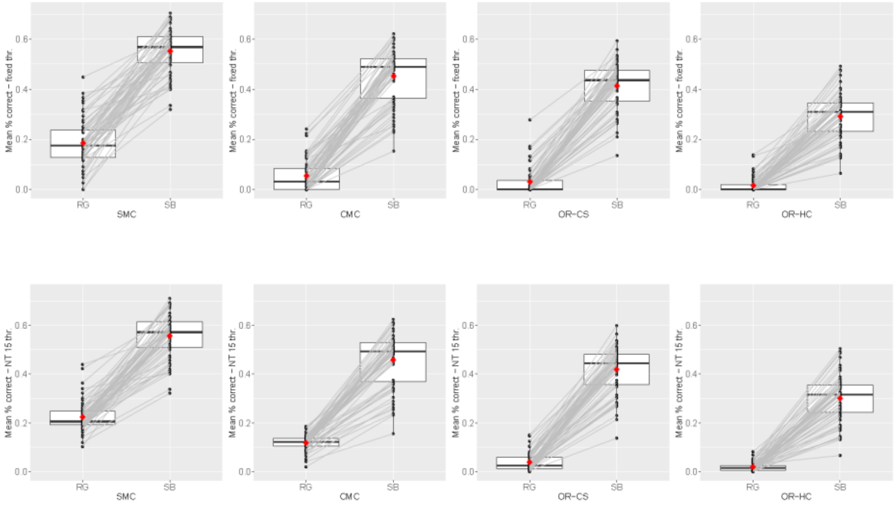

**Figure 1.**
Comparison of percent correct for rapid guessers and solution behavior for each item type and fixed (upper panel) and NT15 (lower panel) thresholds in Mathematics.

*Note*. Points connected with a grey line represent the same jurisdiction.

Occasionally, there were aberrant items for which the percent correct for rapid guessers was higher than for solution-behavior examinees. Averaged across jurisdictions, these inconsistent items were 4.61 % (fixed) and 5.05 % (NT15) of all Mathematics items. The inconsistent items were almost always MC, and more likely to be simple than complex MC.

Non-rapid guessing behaviors in the Mathematics test were aggregated to calculate response time effort for each student's RTE. A student who did not rapid guess on any of the items would get an RTE of 100 %. Then, students' RTE scores were averaged using sampling weights to obtain a weighted RTE score for each jurisdiction. Across jurisdictions, average RTE was very high at 99.3 % with the fixed threshold, and 96.5 % with the NT15 method, mirroring the low levels of rapid guessing reported on Table 1. RTE scores were high for all jurisdictions, since the standard deviations were very small, 0.5 % and 1.5 %, respectively. Figure 2 lists the jurisdictions in decreasing order of weighted mean RTE and both thresholds. The Spearman correlation coefficient for RTE scores under the two thresholds was high, $r_s = .84$, $p < .001$. Macao and Estonia toped the rankings under NT15, while Qatar and Israel had the lowest RTEs; however, overall variability was low and the range was less than 10 %. A map with the weighted mean RTE scores can be seen in Figure M1 in the online Supplementary Material on OSF (https://osf.io/pxq6h/). M2 is the corresponding map for the fixed threshold. Regional patterns are not very clear, except lower scores in the Middle East and eastern Mediterranean countries depicted with darker red colors.

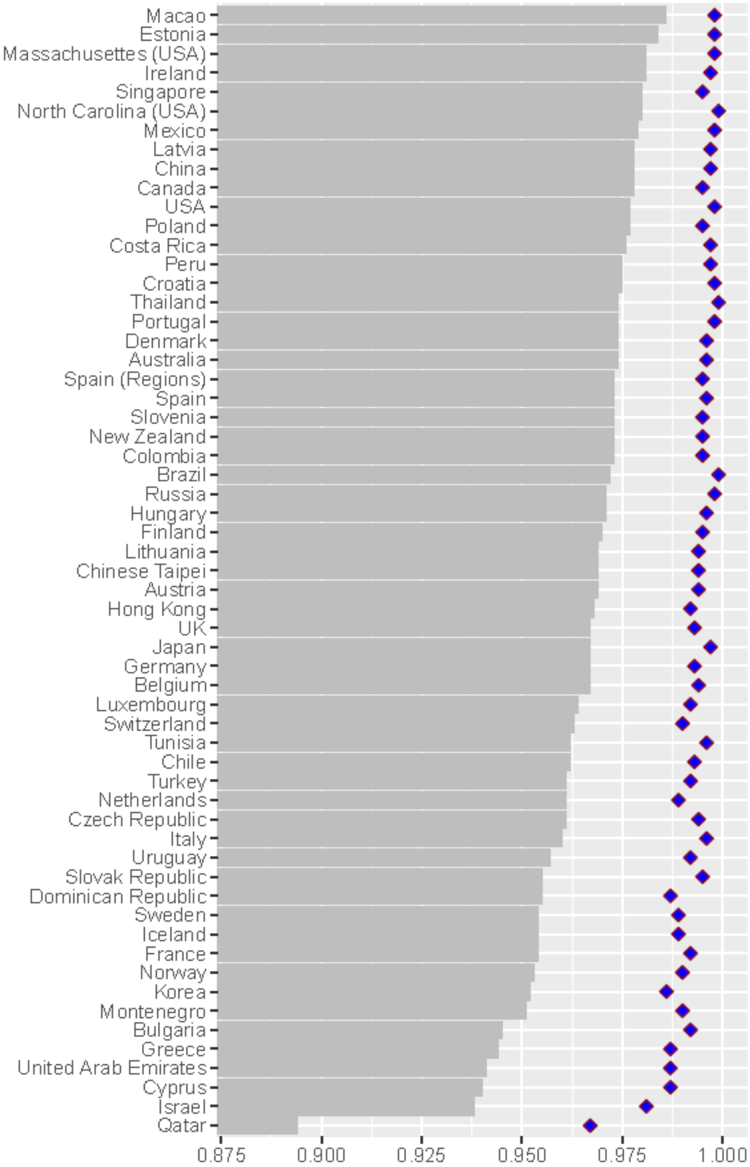

**Figure 2.**
Bar graph for the weighted mean RTE scores by country in Mathematics

*Note*. Bars represent RTEs under NT15 and diamond symbols under the fixed threshold.

Finally, the weighted mean RTE score per jurisdiction was correlated with a number of demographic and personality aggregate indices as can be seen in the online Supplementary Material on OSF (Figure M3). Very few significant results were derived. Under the NT15 method which allowed for larger variability across jurisdictions than the fixed threshold, RTE was positively correlated with country mean score in Mathematics, $r_s = .27$, $p < .05$. GDP and HDI indices were negatively correlated only with the RTE using the fixed method, $r_s = -.33$, $p < .05$, and $r_s = -.27$, $p < .05$. None of the six Hofstede dimensions or the Big Five country scores that were available resulted in significant associations with the RTE indices.

## Summary of Results from the Reading Assessment

The same analyses were conducted with the PISA 2015 Reading data. Similar results to Mathematics were found with respect to the extent of rapid guessing behavior, except with a higher percentage of rapid guessers identified under the fixed threshold (Table 1). The average threshold for NT15 was lower in Reading at 12.64 sec. (*SD* = 1.32). Percentages for each jurisdiction separately are presented in Table A2 in the Appendix. The difference in average percent correct in favor of solution behaviors compared to rapid guessing examinees was even higher in Reading compared to Math: 0.56 to 0.08 for the fixed, and 0.57 to 0.11 for the NT15 thresholds (Table 2). The largest difference was observed on the OR human-coded items (see also the boxplots on Figure 3). Instances of aberrant items where rapid guessers performed better than their peers were fewer in Reading: 2.70 % (fixed) and 2.85 % (NT15), and almost all were of the MC format.

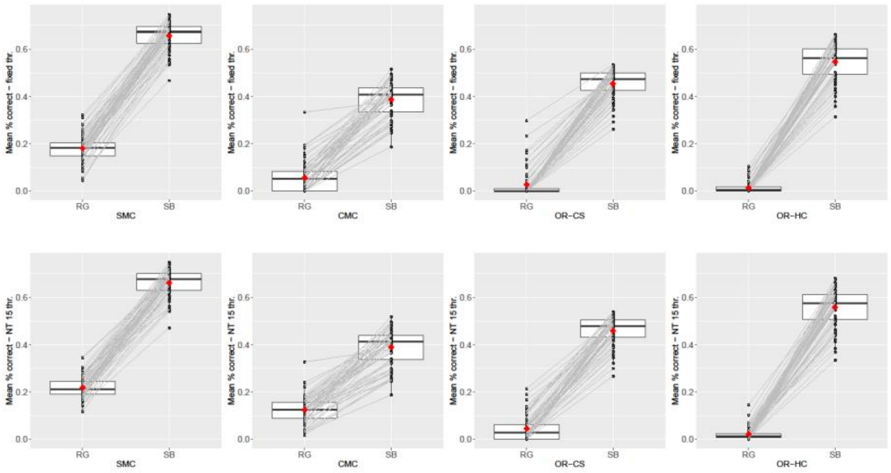

**Figure 3**
Comparison of percent correct for rapid guessers and solution behavior for each item type and fixed (upper panel) and NT15 (lower panel) thresholds in Reading

*Note.* Points connected with a grey line represent the same jurisdiction.

Across jurisdictions, average RTE was very high for Reading at 98.6 % (*sd* = 0.9 %) with the fixed threshold, and 96.7 % (*sd* = 1.6 %) with NT15 (Figure 4). The Spearman correlation coefficient for RTE scores under the two thresholds was high, $r_s$ = .88, *p* < .001. Massachusetts, Estonia, and Mexico had the highest NT15 RTEs, while Qatar and Cyprus had the lowest; however, the range was less than 9 %. Regional patterns indicated lower average RTE scores in the Middle East and eastern Mediterranean countries as in Mathematics (Figure 4 and maps R1 and R2 in the online Supplementary Material on OSF https://osf.io/pxq6h/). Average RTE under the NT15 method was positively correlated with country mean score in Reading, $r_s$ = .37, *p* < .01 (Figure R3, online Supplementary Material). No other correlations were found with demographic or personality country-level variables.

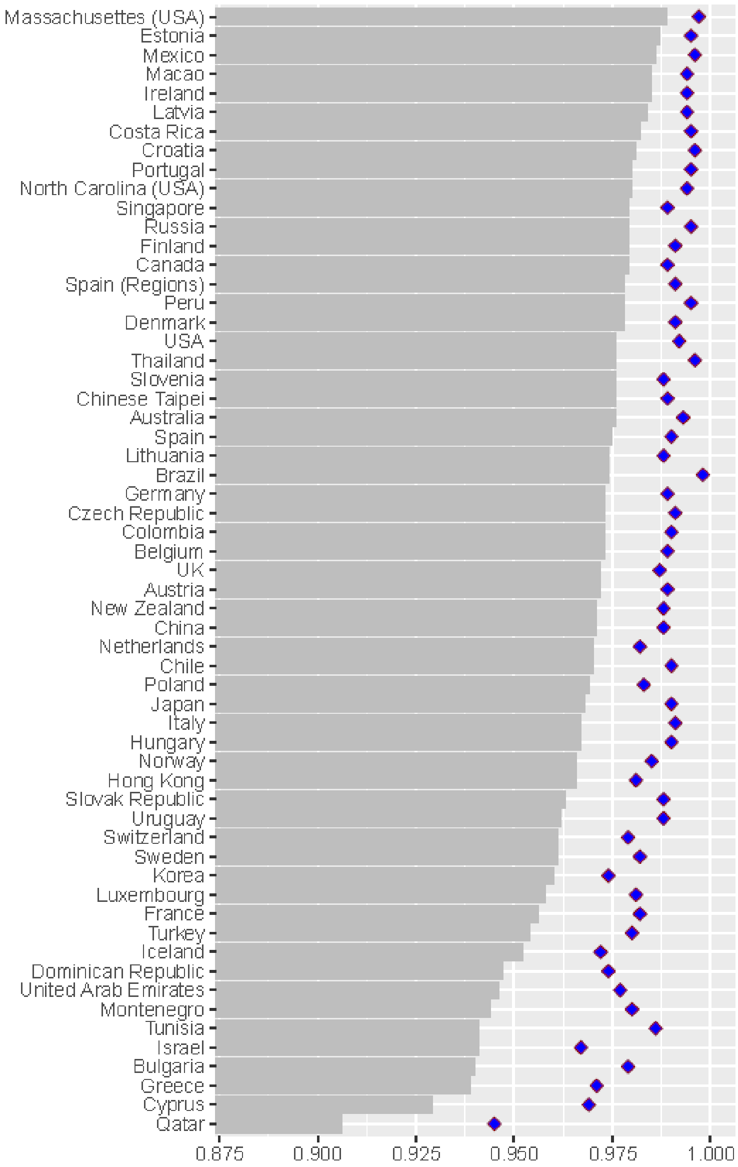

**Figure 4**
Bar graph for the weighted mean RTE scores by country in Reading

*Note*. Bars represent RTEs under NT15 and diamond symbols under the fixed threshold

## Discussion

The current descriptive study portrayed the extent of rapid guessing behavior in the PISA 2015 computerized Mathematics assessment. The focus was on two techniques of identifying rapid guessing based on item response times and constructing aggregate measures at the jurisdiction level; results included comparisons by types of test items. Overall, the proportion of rapid guessers on each item was found to be small on average, about 3 %, according to the normative approach which classified as rapid guessers the examinees who respond faster than the 15 % of the average item response time in the jurisdiction. With the more conservative "fixed" 5-second threshold, that figure was much smaller than 1 %, probably a substantial under-identification; items administered in PISA have sometimes verbose stems and scenarios, and/or accompanied by visual stimuli. We concur with other researchers that a more liberal threshold like NT15 is more informative in capturing rapid guessers (Soland, et al., 2021; Wise, 2019; Wise & Kuhfeld, 2021). Of course, any threshold method that dichotomizes a sample runs the danger of misclassifying examinees. The NT15 may be a reasonable procedure since it is item-specific, provides a threshold for all items, and is moderately liberal for balancing false-positive and false negatives (Lindner, et al., 2019).

The classification of examinees based on rapid guessing behavior, and subsequently the aggregate measure of response-time effort, was validated: average performance for rapid guessers was on average lower than for test-takers engaged in solution behavior (Wise, 2017). This was true in aggregate comparisons for all item types, in all 59 jurisdictions, and in both Mathematics and Reading – more pronounced in the latter. The rare instances where few items were inconsistent, with rapid guessers demonstrating higher accuracy than non-rapid-guessers, concerned almost exclusively selected-response items. This is not unexpected when e.g., the number of rapid guessers is very low and the proportion correct estimate for that group is volatile; and/or the correct option is B or C which are often endorsed by examinees who guess (Attali & Bar-Hillel, 2003; Michaelides, et al., 2020; Wise, 2017).

A novel finding of this study was that there were similar percentages of rapid guessing in simple and complex MC items, and in OR computer-scored items. Rapid guessing was about twice as much in OR human-coded items, which typically require responses to be produced vis-à-vis selected by the examinees. The proportion correct was also lower on average on OR than on MC items, both for rapid guessers and solution-behavior examinees (Figures 1 and 3). In the context of an assessment like PISA where the stakes and motivation are not high for the examinees, these findings are consistent with Krosnick's (1991) predictions about satisficing in surveys: higher task difficulty and demands increase the probability of engaging in less than optimal behavior; "if a task is especially difficult, the combination of low ability and low motivation should powerfully enhance satisficing" (p.225). Successful performance was on average lower especially on Mathematics OR human-coded items (in part because of higher rapid guessing or rapid omits). Research efforts to improve item construction e.g. with novel stimuli and formats, can be directed to item types that are more likely to elicit rapid guessing behavior. Computerized items requiring production rather than selection of a response, and higher cognitive load may be enhanced with digital interactive features to foster engagement.

Another aim of this study was to present cross-country differences in rapid guessing rates, after aggregating the item-level student's behavior to a weighted response time effort for the entire test. As expected, the average RTE for all jurisdictions was very high, 96.5 % (or 99.3 % under the fixed threshold), and with very low variability since all countries had an average RTE of 93 % or higher (with just one exception). In practice this means that students on average rapid guess on about one out of 20 items only, although it should be pointed out that the majority of students had an RTE of 100 %. A common criticism for international large-scale assessments is that the low level of test-taking effort is a primary reason when country performance is low – a criticism that is not empirically supported (Baumert & Demmrich, 2001; Hopfenbeck & Kjærnsli, 2016). The evidence from the current investigation does not support this criticism either, since the extent of rapid guessing is generally small, and relatively homogeneous across most jurisdictions. This empirical evaluation implies that comparative inferences from the PISA assessments are not invalidated from differential RTE.

A number of country-level socio-cultural and psychological characteristics were examined as potential correlates for average RTE. No clear regional differences were observed, except perhaps higher average rapid guessing in countries in the Middle East and the Balkans. A potential next step would be to see if the differential RTE in these countries is consequential for their overall standing in PISA. If so, inferences about the comparability of average country scores may be inaccurate (Rios, 2021). However, given that the variability is small, the expectation is that this would not significantly impact country rankings. In a study with a few countries, Guo and Ercikan (2020) found that the ranking did not change after filtering out examinees with lower RTE scores.

At the country-level, RTE was positively correlated with the country's PISA mean score. A similar trend was reported by Rios and Guo (2020) for a small number of countries, and by Debeer et al. (2014) where countries with lower PISA reading scores had larger decrease in effort throughout the test. Reliable associations of average RTE with other economic and socio-psychological aggregate indices were not found. It is likely that characteristics like conscientiousness (Freund & Holling, 2011), or perceptions of hierarchical social structures may operate at the individual, and not at the aggregate level. Future research could examine these characteristics when measured at the individual student level within a multilevel analysis framework.

The change of the PISA administration from paper-and-pencil to computerized format has numerous implications and potential enhancements for the testing program, specifically for test-taking effort. First, multiple types of items, with novel formats can make the testing experience more interesting and engaging for the test-takers. With adaptive testing, the test-takers are presented with items that have difficulty levels that match their proficiency level and are appropriately challenging, neither too easy nor too difficult. With increasing number of countries in PISA administering the computerized format, recording of item response times is easily accomplished. Thus, it becomes feasible to monitor test-taking effort via the response time proxy. Procedural checks could signal the extent of rapid guessing after the assessment to detect aberrant response time patterns that could invalidate examinee scores. Importantly, automated indicators of response times, such as the ones presented in this study, may be used for increasing test-taking effort with customized interventions on-the-fly, when an examinee is responding rapidly to items, e.g., by notifying a proctor (Wise, Kuhfeld & Soland, 2019) or by providing warnings on the screen nudging towards more engagement with the test material (Wise, Bhola & Yang, 2006). It is also possible to track rapid guessing across PISA cycles and monitor trends in test-taking behavior over time.

Pohl, Ulitzsch, and von Davier (2021) suggested that "test-taking behavior is not a nuisance factor that may confound measurement, but an aspect that provides important information on how examinees approach tasks, which is relevant for real-life outcomes" (p.338). A student's score could be alternatively construed as a composite of response accuracy and other valued behaviors such as speed during the test (rapid guessing, varying speed across the test), and response propensity (item non-response, quitting behavior, revisits etc.) Such a composite could be more predictive of future outcomes like educational attainment or performance in employment settings. Determining the weights in this type of a composite remains an open question for future research.

A method relying on response time to characterize effortless behavior overcomes the drawbacks of self-reports of examinee test-taking effort (Eklöf, 2010; Wise, 2015). The dichotomy of into rapid guessing or solution behavior constitutes a simplistic description of effort, since one could argue that more time on task does not necessarily imply more effort, or that a rapid response may actually reflect efficient test-taking performance based on strategic choices by attentive examinees. Methodological improvements are needed in this area since process variables recently made available by computerized platforms may enrich our understanding of test-taking behavior. Initial empirical studies on alternative variables like type or number of actions (Goldhammer, Naumann, Rölke, Stelter, & Tóth, 2017; Greiff, Niepel, Scherer, & Martin, 2016; Ivanova, Michaelides, & Eklöf, 2020) or eye-tracking measures (Koutsogiorgi & Michaelides, 2022) extend the range of observed behaviors that reflect examinee interactions with test items. Future work would benefit from theoretical formulations that encompass and integrate the complexities of test-taking behavior.

## Declarations

**Availability of data and materials.** Data are available on the PISA 2015 website. Analysis code files and an online Supplementary Material file with figures are available at OSF: https://osf.io/pxq6h/

**Competing interests.** The authors declare no competing interests.

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

**Authors' contributions.** MPM and MI jointly devised the project, wrote the Introduction and Methods, and designed the analysis. MPM drafted the Results, and Discussion. MI reviewed and edited the entire manuscript.

**Acknowledgments.** The authors would like to thank Andreas Economides, and Demetris Avraam for assistance with programming, and Hanna Eklöf and Denise Reis Costa for helpful discussions.

**List of abbreviations**

CS: Computer-scored

HC: Human-coded

MC: Multiple-choice

NT: Normative threshold

OECD: Organization for Economic Co-operation and Development

OR: Open-response

PIAAC: Programme for the International Assessment of Adult Competencies

PISA: Programme for International Student Assessment

RTE: Response Time Effort

SD: standard deviation

TIMSS: Trends in International Mathematics and Science Study

## References

Attali, Y., & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement*, *40*(2), 109-128.

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.

Barry, C. L., & Finney, S. J. (2016). Modeling change in effort across a low-stakes testing session: A latent growth curve modeling approach. *Applied Measurement in Education, 29*(1), 46-64.

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing, 10*(4), 342-363.

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.

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.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*(1), 55-77.

Eklöf, H. (2010). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice, 17*(4), 345-356.

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., & Knekta, E. (2017). Using large-scale educational data to test motivation theories: A synthesis of findings from Swedish studies on test-taking motivation. *International Journal of Quantitative Research in Education, 4*(1-2), 52-71.

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.

Freund, P. A., & Holling, H. (2011). Who wants to take an intelligence test? Personality and achievement motivation in the context of ability testing. *Personality and Individual Differences*, *50*(5), 723-728.

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

Goldhammer, F., Martens, T., & Lüdtke, O. (2017). Conditioning factors of test-taking engagement in PIAAC: An exploratory IRT modelling approach considering person and item characteristics. *Large-Scale Assessments in Education*, *5*(18), 1-25. https://doi.org/10.1186/s40536-017-0051-9

Goldhammer, F., Naumann, J., Rölke, H., Stelter, A., & Tóth, K. (2017). Relating product data to process data from computer-based competency assessment. In D. Leutner, J. Fleischer, J. Grünkorn, & E. Klieme (Eds.), *Competence assessment in education: Research, models and instruments* (pp. 407–425). Springer.

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.

Greiff, S., Niepel, C., Scherer, R., & Martin, R. (2016). Understanding students' performance in a computer-based assessment of complex problem solving: An analysis of behavioral data from computer-generated log files. *Computers in Human Behavior, 61*, 36–46.

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

Guo, H., Rios, J. A., Haberman, S., Liu, O. L., Wang, J., & Paek, I. (2016). A new procedure for detection of students' rapid guessing responses using response time. *Applied Measurement in Education, 29*, 173–183. doi:10.1080/08957347.2016.1171766

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

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.

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.

Kroehne, U., Deribo, T., & Goldhammer, F. (2020). Rapid guessing rates across administration mode and test setting. *Psychological Test and Assessment Modeling*, *62*(2), 147-177.

Krosnick, J. A. (1991). Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied cognitive psychology, 5*(3), 213-236.

Koutsogiorgi, C., & Michaelides. M. P. (2022). Response tendencies due to item wording using eye tracking methodology. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-021-01719-x

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*, 1533.

McCrae, R. R., & Costa, P. T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology, 52*(1), 81-90.

McCrae, R. R., Terracciano, A., & Personality Profiles of Cultures Project. (2005). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology, 89*(3), 407–425.

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.

Organization for Economic Co-operation and Development. (2015). *The ABC of gender equality in education: Aptitude, behaviour, confidence.* OECD Publishing. http://dx.doi.org/10.1787/9789264229945-en

Organization for Economic Co-operation and Development. (2016). PISA 2015 Results Excellence And Equity In Education (Vol. I). OECD Publishing. https://doi.org/10.1787/9789264266490-en

Organization for Economic Co-operation and Development. (2017a). *PISA 2015 technical report.* OECD Publishing. https://www.oecd.org/pisa/data/2015-technical-report/

Organization for Economic Co-operation and Development. (2017b). *PISA2015, assessment and analytical framework. Science, Reading, Mathematics, Financial Literacy and Collaborative Problem Solving* (revised edition). OECD Publishing. http://dx.doi.org/10.1787/9789264281820-en

Organization for Economic Co-operation and Development. (2019). *PISA 2018 results: What students know and can do* (Vol. I). OECD Publishing. https://doi.org/10.1787/5f07c754-en

Pohl, S., Ulitzsch, E., & von Davier, M. (2021). Reframing rankings in educational assessments. *Science*, *372*(6540), 338-340.

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*(10), 1-31. https://doi.org/10.1186/s40536-021-00104-6

Rios, J. A. (2021). Is differential noneffortful responding associated with type I error in measurement invariance testing?. *Educational and Psychological Measurement, 81*(5), 957-979.

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.

Rutkowski, D., & Wild, J. (2015). Stakes matter: Student motivation and the validity of student assessments for teacher evaluation. *Educational Assessment, 20*(3), 165-179.

Schmitt, D. P., Allik, J., McCrae, R. R., & Benet-Martínez, V. (2007). The geographic distribution of Big Five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology, 38*(2), 173–212.

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

Setzer, J. C., Wise, S. L., van den Heuvel, J. R., & Ling, G. (2013). An investigation of examinee test-taking effort on a large-scale assessment. *Applied Measurement in Education, 26*(1), 34-49.

Shin, H. J., Kerzabi, E., Joo, S.-H., Robin, F., & Yamamoto, K. (2020). Comparability of response time scales in PISA. *Psychological Test and Assessment Modeling, 62*(1), 107-135.

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.

Soland, J., Kuhfeld, M., & Rios, J. (2021). Comparing different response time threshold setting methods to detect low effort on a large-scale assessment. *Large-Scale Assessments in Education,* 9(8), 1-21. https://doi.org/10.1186/s40536-021-00100-w

Ulitzsch, E., Penk, C., von Davier, M., & Pohl, S. (2021). Model meets reality: Validating a new behavioral measure for test-taking effort. *Educational Assessment, 26*(2), 104-124.

Ulitzsch, E., Pohl, S., Khorramdel, L., Kroehne, U., & von Davier, M. (2022). A response-time-based latent response mixture model for identifying and modeling careless and insufficient effort responding in survey data. *Psychometrika, 87*(2), 593-619.

Ulitzsch, E., von Davier, M., & Pohl, S. (2020). Using response times for joint modeling of response and omission behavior. *Multivariate Behavioral Research, 55*(3), 425-453.

van de Vijver, F. J., & He, J. (2014). *Report on social desirability, midpoint and extreme responding in TALIS 2013*. OECD Education Working Papers, No. 107, OECD Publishing, Paris. https://doi.org/10.1787/5jxswcfwt76h-en.

Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *The Journal of General Education, 58*(3), 152-166.

Wise, S. L. (2015). Effort analysis: Individual score validation of achievement test data. *Applied Measurement in Education, 28*(3), 237-252.

Wise, S. L. (2017). Rapid-guessing behavior: Its identification, interpretation, and implications. *Educational Measurement: Issues and Practice, 36*(4), 52–61.

Wise, S. L. (2019). An information-based approach to identifying rapid-guessing thresholds. *Applied Measurement in Education, 32*(4), 325-336.

Wise, S. L., Bhola, D. S., & Yang, S. T. (2006). Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT. *Educational Measurement: Issues and Practice, 25*(2), 21-30.

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

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement, 43*(1), 19-38.

Wise, S. L., & DeMars, C. E. (2009). A clarification of the effects of rapid guessing on coefficient α: A note on Attali's "Reliability of speeded number-right multiple-choice tests". *Applied Psychological Measurement, 33*(6), 488-490.

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.

Wise, S. L., & Kingsbury, G. G. (2016). Modeling student Test-Taking motivation in the context of an adaptive achievement test. *Journal of Educational Measurement, 53*(1), 86-105.

Wise, S. L., Kingsbury, G. G., Thomason, J., & Kong, X. (2004, April). *An investigation of motivation filtering in a statewide achievement testing program.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

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.

Wise, S. L., & Kuhfeld, M. R. (2021). Using retest data to evaluate and improve effort-moderated scoring. *Journal of Educational Measurement*, *58*(1), 130-149.

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.

Wise, S. L., & Ma, L. (2012). *Setting response time thresholds for a CAT item pool: The normative threshold method*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Canada, Vancouver.

Wise, S. L., Ma, L., Kingsbury, G. G., & Hauser, C. (2010). *An investigation of the relationship between time of testing and test-taking effort.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, CO, Denver.

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.

## APPENDIX

**Table A1**
Number and country % of rapid guessers by type of threshold for PISA 2015 Mathematics.

| Country | Sample size | Mean number of RG (s.d.) for fixed thr. | Mean % of RG for fixed threshold | Mean number of RG (s.d.) for NT 15 thr. | Mean % of RG for NT 15 threshold | Mean NT 15 item threshold (s.d.) |
|---|---|---|---|---|---|---|
| Australia | 5924 | 5.09 (5.97) | 0.34 | 42.48 (42.41) | 2.87 | 13845.24 (4847.67) |
| Austria | 2866 | 3.86 (6.3) | 0.53 | 20.6 (23.76) | 2.88 | 14074.2 (4476.12) |
| Belgium | 4088 | 3.58 (4.95) | 0.35 | 24.8 (29.84) | 2.48 | 15229.18 (5440.22) |
| Brazil | 9288 | 1.99 (4.29) | 0.09 | 56.07 (62.6) | 2.65 | 16999.02 (4969.7) |
| Bulgaria | 2423 | 4.41 (5.71) | 0.72 | 30.51 (26.07) | 5.03 | 15881.32 (5230.53) |
| Canada | 8213 | 10.84 (11.68) | 0.52 | 54.04 (62.26) | 2.62 | 14508.4 (5705.12) |
| Chile | 2863 | 3.91 (6.33) | 0.55 | 21.63 (26.02) | 3.07 | 16090.55 (5581.38) |
| Chinese Taipei | 3166 | 4.89 (4.63) | 0.61 | 25.33 (25.97) | 3.15 | 13846.32 (6311.92) |
| Colombia | 4787 | 5.72 (10.84) | 0.48 | 33.22 (33.56) | 2.79 | 18268.19 (6071.73) |
| Costa Rica | 3218 | 2.01 (3.69) | 0.30 | 15.95 (20.56) | 2.30 | 16497.56 (5585.5) |
| Croatia | 2379 | 1.01 (2.58) | 0.17 | 13.6 (18.12) | 2.28 | 14792.94 (4334.33) |
| Cyprus | 2291 | 7.6 (9.23) | 1.31 | 36.2 (29.16) | 6.27 | 13363.73 (4425.04) |
| Czech Republic | 2854 | 2.63 (4.45) | 0.37 | 20.47 (24.47) | 2.90 | 15926.36 (5176.27) |
| Denmark | 3131 | 3.75 (5.25) | 0.50 | 22 (25.57) | 2.97 | 14887.08 (5193.84) |
| Dominican Republic | 2525 | 7.19 (11.38) | 1.15 | 26.05 (26.32) | 4.24 | 17492.62 (5302.39) |
| Estonia | 2314 | 0.75 (1.21) | 0.13 | 8.86 (13.9) | 1.52 | 16398.48 (6492.48) |

| Country | Sample size | Mean number of RG (s.d.) for fixed thr. | Mean % of RG for fixed threshold | Mean number of RG (s.d.) for NT 15 thr. | Mean % of RG for NT 15 threshold | Mean NT 15 item threshold (s.d.) |
|---|---|---|---|---|---|---|
| Finland | 2427 | 2.22 (2.56) | 0.36 | 16.17 (18.25) | 2.68 | 14685.4 (5387.23) |
| France | 2485 | 4.17 (7.16) | 0.66 | 24.88 (27.89) | 4.00 | 15703.11 (5713.31) |
| Germany | 2730 | 3.1 (4.41) | 0.45 | 16.27 (20.79) | 2.41 | 14423.53 (5249.55) |
| Greece | 2266 | 6.28 (7.48) | 1.10 | 27.72 (26.02) | 4.89 | 16548.44 (5473.29) |
| Hong Kong | 2212 | 4.25 (4.29) | 0.75 | 16.95 (17.96) | 3.01 | 13439.15 (6509.72) |
| Hungary | 2291 | 1.52 (2.76) | 0.26 | 14.47 (16.35) | 2.51 | 14649.25 (4720.28) |
| Iceland | 1396 | 3.35 (3.85) | 0.96 | 14.84 (13.15) | 4.28 | 14391.8 (4401.01) |
| Ireland | 3105 | 2.57 (3.76) | 0.32 | 14.94 (20.87) | 1.85 | 15323.14 (5832.69) |
| Israel | 2779 | 10.83 (13.53) | 1.56 | 37.93 (33.44) | 5.52 | 16023.65 (5278.53) |
| Italy | 4730 | 3.69 (5.17) | 0.31 | 36.86 (38.94) | 3.15 | 16406.45 (5247.25) |
| Japan | 2722 | 2.23 (4.05) | 0.32 | 21.56 (27.34) | 3.18 | 15214.13 (7315.62) |
| Korea | 2268 | 7.89 (7.32) | 1.37 | 27.2 (26.68) | 4.75 | 11974.56 (5049.06) |
| Latvia | 1994 | 1.37 (2.24) | 0.27 | 10.54 (12.22) | 2.11 | 15632.16 (5174.03) |
| Lithuania | 2692 | 3.07 (4.06) | 0.45 | 18.94 (21.38) | 2.80 | 14720.55 (4363.28) |
| Luxembourg | 2140 | 3.83 (6.4) | 0.71 | 18.4 (19.32) | 3.43 | 15536.57 (5393.73) |
| Macao | 1834 | 0.86 (1.24) | 0.19 | 6.21 (8.56) | 1.34 | 16388.72 (7422.5) |
| Mexico | 3111 | 1.33 (2.36) | 0.18 | 14.26 (15.63) | 1.90 | 18815.15 (6731.39) |
| Montenegro | 2319 | 5.06 (7.63) | 0.87 | 26.81 (26.22) | 4.68 | 14351.69 (3532.68) |
| Netherlands | 2270 | 5.53 (6.49) | 0.97 | 19.06 (18.02) | 3.38 | 12423.37 (4636.72) |

| Country | Sample size | Mean number of RG (s.d.) for fixed thr. | Mean % of RG for fixed threshold | Mean number of RG (s.d.) for NT 15 thr. | Mean % of RG for NT 15 threshold | Mean NT 15 item threshold (s.d.) |
|---|---|---|---|---|---|---|
| New Zealand | 1866 | 2.02 (3.51) | 0.43 | 11.95 (15.13) | 2.59 | 14510.71 (5340.38) |
| Norway | 2229 | 4.77 (6.41) | 0.85 | 24.88 (25.49) | 4.47 | 15493 (5110.57) |
| Peru | 2836 | 1.94 (4.6) | 0.29 | 16.99 (21.11) | 2.53 | 21066.25 (6520.97) |
| Poland | 2419 | 2.96 (4.3) | 0.47 | 14.35 (17.23) | 2.29 | 16004.51 (5704.44) |
| Portugal | 2980 | 1.36 (2.93) | 0.18 | 19.94 (24.91) | 2.71 | 15625.41 (5402.03) |
| Qatar | 6403 | 48.79 (45.61) | 2.97 | 165.51 (103.48) | 10.17 | 13915.52 (4352.11) |
| Russian Federation | 2463 | 1.48 (2.85) | 0.24 | 17.58 (18.83) | 2.90 | 17317.05 (5373.57) |
| Singapore | 2504 | 2.75 (3.13) | 0.43 | 12.94 (14.1) | 2.03 | 14414.16 (6511.88) |
| Slovak Republic | 2704 | 2.32 (3.66) | 0.35 | 25.6 (22.53) | 3.86 | 15265.44 (4423.67) |
| Slovenia | 2708 | 3.89 (5.9) | 0.58 | 22.23 (25.55) | 3.34 | 13954.73 (4407.48) |
| Spain | 2742 | 2.46 (5.05) | 0.35 | 16.69 (21.1) | 2.43 | 16529.44 (5550.48) |
| Sweden | 2218 | 4.74 (6.57) | 0.88 | 23.01 (25.61) | 4.30 | 16511.56 (5784.38) |
| Switzerland | 3159 | 7.1 (8.04) | 0.87 | 28.2 (26.72) | 3.47 | 15192.49 (5433.01) |
| Thailand | 3374 | 0.96 (1.65) | 0.11 | 22.28 (19.72) | 2.66 | 17252.02 (5800.91) |
| United Arab Emirates | 5746 | 17.28 (16.17) | 1.19 | 80.81 (55.41) | 5.61 | 15767.12 (5927.06) |
| Tunisia | 2097 | 1.89 (3.27) | 0.38 | 18.28 (19.01) | 3.66 | 17339.33 (6015.09) |
| Turkey | 2406 | 4.05 (5.17) | 0.66 | 23.14 (20.34) | 3.81 | 14978.35 (4131.8) |
| United Kingdom | 5806 | 8.73 (10.53) | 0.59 | 43.81 (48.25) | 2.99 | 13316.77 (4763.59) |
| United States | 2366 | 1.05 (1.5) | 0.18 | 12.01 (13.52) | 2.04 | 14864.37 (5150.45) |

| Country | Sample size | Mean number of RG (s.d.) for fixed thr. | Mean % of RG for fixed threshold | Mean number of RG (s.d.) for NT 15 thr. | Mean % of RG for NT 15 threshold | Mean NT 15 item threshold (s.d.) |
|---|---|---|---|---|---|---|
| Uruguay | 2466 | 4.52 (9.13) | 0.75 | 24.12 (25.73) | 4.03 | 16174.64 (4948.37) |
| B-S-J-G (China) | 4020 | 3.32 (4.29) | 0.32 | 22.04 (25.95) | 2.16 | 15727.04 (7088.45) |
| Spain (Regions) | 13223 | 14.14 (24.45) | 0.42 | 86.63 (104.91) | 2.62 | 16257.29 (5473.56) |
| Massachusettes (USA) | 711 | 0.46 (0.78) | 0.26 | 3.35 (3.41) | 1.95 | 14860.46 (5732.31) |
| North Carolina (USA) | 783 | 0.22 (0.57) | 0.11 | 3.73 (4.17) | 1.92 | 13969.07 (4636.62) |

**Table A2.**
Number and country % of rapid guessers by type of threshold for PISA 2015 Reading.

| Country | Sample size | Mean number of RG (s.d.) for fixed thr. | Mean % of RG for fixed threshold | Mean number of RG (s.d.) for NT 15 thr. | Mean % of RG for NT 15 threshold | Mean NT 15 item threshold (s.d.) |
|---|---|---|---|---|---|---|
| Australia | 5950 | 12.17 (11.1) | 0.82 | 42.75 (38.74) | 2.88 | 11541.77 (5593.13) |
| Austria | 2889 | 7.99 (9.22) | 1.11 | 20.01 (19.08) | 2.77 | 11326.67 (5783.29) |
| Belgium | 4096 | 9.11 (9.61) | 0.93 | 23.16 (21.34) | 2.37 | 12102.35 (5820.95) |
| Brazil | 9317 | 4.03 (4.86) | 0.21 | 50.97 (57.26) | 2.61 | 14842.85 (6249.23) |
| Bulgaria | 2417 | 11.45 (12.75) | 1.93 | 33.26 (30.33) | 5.60 | 13003.29 (5970.49) |
| Canada | 8214 | 21.63 (19.01) | 1.05 | 46.33 (40.3) | 2.24 | 11536.89 (5961.8) |
| Chile | 2862 | 5.29 (5.48) | 0.77 | 15.93 (16.51) | 2.31 | 13071.67 (6455.35) |
| Chinese Taipei | 3172 | 8.7 (7.54) | 1.08 | 20.27 (18.38) | 2.51 | 11870.79 (6653.52) |
| Colombia | 4810 | 9.48 (8.69) | 0.82 | 27.04 (26.74) | 2.33 | 14701.39 (7127.77) |
| Costa Rica | 3195 | 3.79 (4.39) | 0.60 | 13.06 (13.77) | 2.09 | 13762.97 (6772.75) |
| Croatia | 2401 | 2.27 (2.84) | 0.37 | 10.47 (13.65) | 1.72 | 11812.2 (5570.88) |
| Cyprus | 2257 | 17.27 (17.03) | 3.10 | 40.19 (33.25) | 7.21 | 11675.54 (5551.25) |
| Czech Republic | 2882 | 5.19 (5.54) | 0.74 | 15.58 (15.3) | 2.23 | 12187.66 (5946.44) |
| Denmark | 3128 | 9.16 (10.08) | 1.29 | 21.36 (19.26) | 3.00 | 12131.87 (5820.66) |
| Dominican Republic | 2535 | 14.04 (20.17) | 2.39 | 29.89 (32.29) | 5.13 | 15195.15 (6842.8) |
| Estonia | 2278 | 2.51 (4.58) | 0.44 | 6.92 (9.28) | 1.21 | 12351.7 (6206.81) |
| Finland | 2436 | 4.53 (4.25) | 0.74 | 11.43 (10.85) | 1.87 | 11591.77 (5868.99) |

| Country | Sample size | Mean number of RG (s.d.) for fixed thr. | Mean % of RG for fixed threshold | Mean number of RG (s.d.) for NT 15 thr. | Mean % of RG for NT 15 threshold | Mean NT 15 item threshold (s.d.) |
|---|---|---|---|---|---|---|
| France | 2492 | 9.8 (11.24) | 1.57 | 24.19 (24.71) | 3.88 | 12485.32 (6339.44) |
| Germany | 2743 | 5.74 (5.99) | 0.87 | 14.35 (14.93) | 2.16 | 12166.94 (6466.31) |
| Greece | 2278 | 13.53 (11.86) | 2.37 | 28.96 (23.03) | 5.06 | 13826.8 (7778.82) |
| Hong Kong | 2185 | 9.53 (10.13) | 1.72 | 17.71 (15.33) | 3.20 | 12094.05 (7511.44) |
| Hungary | 2331 | 4.79 (7.11) | 0.82 | 16.07 (18.37) | 2.76 | 11869.88 (5446.7) |
| Iceland | 1374 | 8.72 (11.29) | 2.56 | 15.61 (15.28) | 4.58 | 11623.66 (5998.21) |
| Ireland | 3088 | 4.73 (4.66) | 0.59 | 11.81 (12.2) | 1.48 | 12867.1 (6841.92) |
| Israel | 2771 | 20.81 (15.55) | 3.20 | 37.45 (24.67) | 5.72 | 12583.6 (6307.57) |
| Italy | 4754 | 7.91 (7.75) | 0.68 | 30.12 (28.6) | 2.58 | 12855.4 (6360.97) |
| Japan | 2692 | 6.15 (6.23) | 0.92 | 20.82 (22.67) | 3.10 | 13966.31 (8202.74) |
| Korea | 2286 | 14.94 (14.31) | 2.56 | 22.78 (19.04) | 3.91 | 9662.49 (4912.94) |
| Latvia | 2000 | 3.12 (3.29) | 0.62 | 7.89 (7.48) | 1.57 | 12481.47 (6300.07) |
| Lithuania | 2659 | 6.84 (6.09) | 1.01 | 17.2 (15.58) | 2.55 | 11712.86 (5675.16) |
| Luxembourg | 2193 | 9.92 (9.72) | 1.82 | 21.8 (18.56) | 4.00 | 12614.4 (6868.59) |
| Macao | 1836 | 2.35 (2.89) | 0.52 | 6.34 (7.15) | 1.38 | 13503.48 (7664.3) |
| Mexico | 3044 | 2.27 (2.71) | 0.31 | 9.24 (11.32) | 1.28 | 14504.54 (7041.92) |
| Montenegro | 2302 | 10.73 (10.93) | 1.93 | 30.73 (25.96) | 5.52 | 12118.67 (5314.38) |
| Netherlands | 2261 | 9.29 (10.32) | 1.69 | 15.52 (14.81) | 2.83 | 10327.65 (4898.29) |
| New Zealand | 1842 | 4.84 (5.18) | 1.05 | 12.51 (12.16) | 2.71 | 12047.38 (5995.78) |

| Country | Sample size | Mean number of RG (s.d.) for fixed thr. | Mean % of RG for fixed threshold | Mean number of RG (s.d.) for NT 15 thr. | Mean % of RG for NT 15 threshold | Mean NT 15 item threshold (s.d.) |
|---|---|---|---|---|---|---|
| Norway | 2235 | 7.72 (7.13) | 1.38 | 18.29 (15.98) | 3.27 | 11878.01 (5918.85) |
| Peru | 2832 | 2.69 (3.13) | 0.43 | 13.12 (13.9) | 2.11 | 17536.66 (7645.74) |
| Poland | 2406 | 10.13 (15.47) | 1.63 | 19.14 (23.18) | 3.08 | 12450.67 (6074.8) |
| Portugal | 3025 | 3.59 (5.48) | 0.48 | 16.98 (20.33) | 2.26 | 12856.82 (6361.2) |
| Qatar | 6419 | 80.86 (63.94) | 5.04 | 139.25 (100.78) | 8.66 | 11366.88 (5564.22) |
| Russian Federation | 2446 | 3.24 (3.64) | 0.55 | 13.69 (14.14) | 2.31 | 13788.65 (6881.75) |
| Singapore | 2513 | 6.96 (7.35) | 1.09 | 14.43 (13.39) | 2.26 | 12385.65 (6650.2) |
| Slovak Republic | 2659 | 6.88 (7.03) | 1.07 | 22.46 (18.74) | 3.49 | 12054.49 (5520.31) |
| Slovenia | 2703 | 10.03 (13.24) | 1.52 | 19.89 (20.27) | 3.02 | 10694.58 (5007.43) |
| Spain | 2762 | 6.02 (5.94) | 0.90 | 14.95 (14.4) | 2.21 | 12839.42 (6586.27) |
| Sweden | 2277 | 8.92 (8.63) | 1.61 | 19.84 (17.94) | 3.55 | 12554.39 (6447.66) |
| Switzerland | 3146 | 15.42 (13.39) | 1.94 | 29.38 (22.78) | 3.70 | 11852.83 (6151.53) |
| Thailand | 3363 | 3.17 (4.38) | 0.39 | 20.01 (22.79) | 2.45 | 14280.84 (6999.75) |
| United Arab Emirates | 5737 | 30.74 (27.48) | 2.16 | 73.89 (56.25) | 5.20 | 13017.74 (6190.2) |
| Tunisia | 2076 | 5.98 (6.12) | 1.28 | 27.2 (25.35) | 5.78 | 15851.17 (7693.62) |
| Turkey | 2422 | 12.78 (13.04) | 2.08 | 29.46 (23.73) | 4.81 | 12983.64 (6238.47) |
| United Kingdom | 5752 | 20.35 (19.98) | 1.39 | 42.92 (38.47) | 2.95 | 11186.32 (5476.71) |
| United States | 2346 | 4.59 (5.32) | 0.80 | 13.84 (13.05) | 2.40 | 12390.86 (6178.38) |
| Uruguay | 2434 | 6.04 (7.46) | 1.07 | 19.67 (19.52) | 3.48 | 13835.24 (6528.82) |

| Country | Sample size | Mean number of RG (s.d.) for fixed thr. | Mean % of RG for fixed threshold | Mean number of RG (s.d.) for NT 15 thr. | Mean % of RG for NT 15 threshold | Mean NT 15 item threshold (s.d.) |
|---|---|---|---|---|---|---|
| B-S-J-G (China) | 4025 | 10.21 (9.06) | 1.00 | 24.3 (22.23) | 2.37 | 13188.35 (7744.37) |
| Spain (Regions) | 13236 | 26.54 (24.82) | 0.82 | 66.02 (62.68) | 2.02 | 12693.52 (6518.16) |
| Massachusettes (USA) | 677 | 0.5 (0.92) | 0.31 | 1.87 (2.2) | 1.16 | 12176.1 (6421.71) |
| North Carolina (USA) | 822 | 1.12 (1.51) | 0.58 | 3.99 (4.39) | 2.05 | 12074.38 (5920.09) |