Psychological Test and Assessment Modeling, Volume 62, 2020 (2), 179-231

Examining gender DIF and gender differences in the PISA 2018 reading literacy scale: A partial invariance approach

Lale Khorramdel[1], *Artur Pokropek*[2], *Seang-Hwane Joo*[3], *Irwin Kirsch*[3] & *Laura Halderman*[3]

Abstract

Gender differences in reading performance in international large-scale assessments (ILSAs) are regularly observed across countries and assessments and over time. This paper aims to evaluate different sources of gender differences in PISA 2018. First, we evaluate whether gender differences might be related to gender-specific differential item functioning (DIF). For analyzing DIF in the complex settings of ILSAs, a multiple-group concurrent calibration based on the two-parameter logistic model (2PLM) and generalized partial credit model (GPCM) with partial invariance assumption is used. Second, we examine the diagnostic value of the reading literacy subscales (text sources, text formats, cognitive processes) as well as students' attitude towards reading through use of multidimensional item response theory (MIRT) models, linear regression, and other exploratory analysis. Results show no strong DIF effects for gender in PISA 2018 and that no additional value is provided by the reading literacy subscales. We show that gender differences might, in part, be related to reading attitudes, at least in some country-by-language groups.

Keywords: differential item functioning, gender differences, item response theory, measurement invariance, PISA

[1] *Correspondence concerning this article should be addressed to:* Lale Khorramdel, PhD, Center for Advanced Assessments (CAA), National Board of Medical Examiners, 3750 Market Street, Philadelphia, PA 19104, USA; email: lale.khorramdelameri@gmail.com

[2] Institute of Philosophy and Sociology, Polish Academy of Sciences (IFiS PAN), Warsaw, Poland

[3] [3]Educational Testing Service (ETS), Princeton, USA

Introduction

Gender differences in educational an reading assessments

Gender differences in educational assessments that use standardized tests have been the subject of numerous studies for at least half a century (Borgonovi & Grieff, 2020; Buchmann et al., 2008; Maccoby & Jacklin, 1978). Strong evidence exists for two patterns: boys having higher test scores in mathematics, and girls having higher test scores in reading and literacy (Baker & Jones, 1993; Beller & Gafni, 1996; Breda & Napp, 2019; Gallagher & Kaufman, 2005; Nowell & Hedges, 1998; Rapp & Borgonovi 2019; Reilly et. al., 2019).

In general, the gender gap in mathematics has been shown to be rather small in the early stages of education but to increase in later educational stages (Willingham & Cole, 2013) and early adulthood (Borgonovi et al., 2018). While past research has documented a stable and persistent gender gap in mathematics (Nowell & Hedges, 1998), newer studies suggest this gap has been closing since the beginning of 21st century, at least in industrialized countries (Hyde & Mertz, 2009; Lindberg et al., 2010; Organisation for Economic Co-operation and Development [OECD], 2019a).

Differences between boys and girls in reading proficiency are much higher than in mathematics (Lietz, 2006; Stoet & Geary, 2013). In PIRLS 2016, girls in fourth grade showed higher average achievement than boys; the average effect size difference between mean scale scores across countries (using Cohen's *d;* Cohen, 1977) was 0.19. Out of 50 countries, only two had statistically non-significant differences: Macao and Portugal. Among industrialized countries, Finland, Norway and Australia showed the largest gender gaps ($d = 0.22$ for Finland, and $d = 0.21$ for Norway and Australia) (Mullis et al., 2017). Similar to the gap in mathematics, the gap in reading showed an upward trend in later school years (Buchmann et al., 2008; Nowell & Hedges, 1998; Willingham & Cole, 2013). In the Programme for International Student Assessment (PISA), which assesses the achievement of 15-year-old students, gender gaps in reading have been quite large and have appeared consistently across different assessment cycles. In PISA 2018, evidence indicates that the gender gap for reading literacy on average shows an effect size of 0.30 (Cohen's *d*), although ranging as high as 0.52 for Finland, 0.48 for Israel, 0.47 for Norway, and 0.42 for Greece (OECD, 2019a). However, differences disappear or become negligible in surveys that focus on adult populations, as supported by evidence from the International Adult Literacy Survey (IALS) conducted between 1994 and 1998 (OECD & Statistics Canada, 2005) and the Programme for the International Assessment of Adult Competencies (PIAAC) 2012 (Borgonovi et al., 2018). Similar patterns of overall differences between boys and girls in reading proficiency can be found in national large-scale assessments, including the National Assessment of Educational Progress (NAEP) in the United States (Mau & Lynn, 2000) and the Australian National Assessment Program Literacy and Numeracy (NAPLAN; Watson et al., 2016).

Studies other than large-scale assessments (using researcher developed reading tests) show much smaller overall gender differences in reading. Hyde and Linn (1988) reported an effect size of 0.09 (Cohen's *d*) for 6- to 10-year-old students in their meta-analysis

including 18 studies of reading comprehension, 0.02 for 11- to 18-year-old examinees, and -0.03 for 19- to 25-year-old examinees for reading achievement across 18 studies. Small to negligible gender differences (d = 0.11) in reading were also found in 16 German studies on elementary school students (Mücke, 2009).

Possible sources of the gender gap in reading assessments

Three sets of nonexclusive hypotheses are used to explain why females tend to outperform males in reading at school: (1) sociocultural, (2) biocognitive, and (3) test-taking behavior.

Sociocultural hypotheses focus on the importance of cultural factors that reward females for reading. The main argument is that through the socialization of girls and boys within their families and schools, traditional gender stereotypes and norms influence students' motivation for reading and their perception of their abilities. Although the causal claims about the mechanisms still need more research, the support in the outcome variables seems strong. Girls have reported higher intrinsic reading motivation on average than boys (McGeown et al., 2012; McKenna et al., 2012), displayed more positive attitudes towards reading, and showed higher frequencies of reading behavior (Kennedy, 2008; Logan & Johnston, 2009). Moreover, it was found that the lower levels of reading ability for boys are related to the self-concept and subjective value in reading (Petscher, 2010).

Biocognitive hypotheses focus on biological predispositions of females toward reading. The research has shown no substantial differences between males and females in general abilities (Mackintosh, 1996) and fluid reasoning (Camarata & Woodcock, 2006; Kaufman & Horn, 1996). Some studies suggest that the level of specific cognitive abilities could be different among genders and follow different developmental trajectories. Females have shown a consistent advantage in processing speed (Keith et al., 2008) and appear to develop language skills earlier than males on average (Bornstein et al., 2004). Furthermore, neuroimaging studies have suggested that males and females display different patterns of functional activation during reading (Logan & Johnston, 2009).

Finally, *hypotheses related to test-taking behavior* focus on the differential performance of boys and girls on cognitive tests. These hypotheses are not independent of the other presented ones and assume that test-taking behavior could be an important mediator and moderator of sociocultural and biological influences (Borgonovi, 2016; DeMars et. al. 2013; Penk, & Richter, 2017).

The first set of behavior related hypotheses focus on the concept of test-taking motivation. Test-taking motivation determines the extent to which examinees make an effort to accurately represent their knowledge in the content area covered by the test (Wise & DeMars, 2005, p. 2). With low-stakes assessments, there is substantial concern about the level of test-taking motivation. In some studies, little influence from test-taking motivation on test performance has been demonstrated (O'Neil et al., 1995; Wise & Kong, 2005). However, in most empirical studies, researchers have shown a positive (moderate to strong) relationship between test performance and test-taking motivation (Sundre & Kitsantas, 2004; Thelk et al., 2009; Wise & Kong, 2005; Wolf & Smith, 1995). In the

meta-analytic review of several studies concerning the relationship between test-taking motivation and test performance by Wise and DeMars (2005), the average effect size (Cohen's *d*) was 0.59, indicating that motivated students performed more than about half a standard deviation better than students who were not motivated. Most importantly, the level of test-taking motivation varied across gender groups for both math and reading tests (DeMars et al., 2013; Eklöf, 2010).

The second set of test-taking behavior hypotheses deals with the specific content and requirements of reading tests (Willingham & Cole, 2013), which can be analyzed by investigating the features of test items. Previous studies have provided evidence that the gender gap is larger for open-ended questions (Beller & Gafni, 2000), continuous texts (OECD & Statistics Canada, 2005), and more cognitively demanding reading tasks (Lafontaine & Monseur, 2009; Schwabe et al., 2015). Moreover, differences in cognitive processes and aspects of reading (comprehension process, retrieve, straightforward inferences, interpret and integrate) and other types of text (fiction versus nonfiction, digital versus printed) were found as well (Solheim & Lundetræ, 2018).

Then, there is the problem of differential item functioning (DIF) across gender groups as a potential source of gender differences. An item displays DIF if test-takers (e.g. students) from different groups (e.g., female and male students) with the same underlying true ability have a different probability of giving a correct response to the item. That is, the item might measure different constructs in different groups (Holland & Wainer, 1993). Gender differences should be examined based on a scale that provides a fair measurement across gender groups. This means, that gender DIF needs to be removed from the scale or, in other words, accounted for.

Aims of this study

In this article, we aim to examine the role of gender DIF, the role of different aspects related to reading (text sources, text formats, cognitive processes), and students' attitudes towards reading as potential sources for the large and consistent gender differences found in PISA. To our knowledge, the present study is the first comprehensive study that focusses on gender DIF for the PISA reading literacy scale on an international level. Because the PISA 2018 reading literacy scale reflects the different aspects we are interested in, it is a great source for such a study. Moreover, the PISA student background questionnaire (BQ) provides variables related to reading attitudes. We examine the following research questions:

1. Can the observed gender differences in reading literacy be explained by a substantial amount of gender DIF?
2. Do the reading literacy subscales' text sources (single, multiple), text formats (continuous, noncontinuous, mixed), and cognitive processes (locate information, understand, evaluate and reflect) provide diagnostic value for understanding gender differences beyond the main score for reading literacy?
3. Can variables related to students' attitudes towards reading explain some of the observed gender differences?

Methods

Data

We used the PISA 2018 main survey data to examine these research questions. PISA is a major international student survey assessing skills of 15-year-old students in the core domains of reading literacy, mathematical literacy, and scientific literacy. PISA has been administered in cycles every three years, starting in 2000. In each cycle, one of the core domains is treated as the major domain, meaning it is administered to all students, while the other domains are considered minor domains and are not administered to all students. Because reading literacy was the major domain in PISA 2018, all students received reading literacy items in addition to either mathematical or scientific literacy items. While the minor domains consist of trend items[4] only, the major domain consists of trend items as well as newly developed items that reflect the updated framework. Moreover, PISA has been administered as a computer-based assessment (CBA) in most participating countries since 2015; only a few countries still were using a paper-based assessment (PBA; note that PBAs include trend items only, even for the major domain, as new items were only developed for the CBA).

In PISA 2018, the reading literacy domain consisted of 346 items which were administered through a multistage adaptive test (MSAT) design: 172 new (computer-based) items,[5] 72 trend items in the CBA, and 102 trend items in the paper-based assessment. Of the 102 paper-based trend items, 72 were identical to the computer-based trend items. However, they were treated as separate items in the analysis because a subset of them received different item parameter estimates for PBA and CBA due to mode effects discovered in the PISA 2015 analysis.[6]

The PISA 2018 data utilized in our study came from 79 countries, which were split into 116 country-by-language groups and 232 country-by-language-by-gender groups for the IRT scaling and DIF analysis. The total sample amounts to N = 619,508 students. We considered the same country-by-language groups as in the operational PISA analyses (to take the effect of multiple languages within a country into account) and excluded *une heure*[7] cases. Moreover, we used senate weights in our analysis, again following the procedure taken in the operational settings. The senate weights are scaled to sum up to 5,000 for each country so that all countries contribute equally to the IRT scaling.

[4] Trend items are items that were administered in previous assessment cycles, and that are used for linking across cycles.

[5] An additional item needed to be excluded due to a coding error.

[6] For more details on the related mode effect study, see OECD (2017a); also please note that we do not include reading fluency items in our analysis.

[7] Une heure refers to a PISA assessment shortened to one hour (as opposed to two hours) for students with special needs.

PISA 2018 reading literacy

Reading literacy in PISA 2018 is defined as "understanding, using, evaluating, reflecting on and engaging with texts in order to achieve one's goals, to develop one's knowledge and potential and to participate in society" (OECD, 2019b). The reading process is assumed to be influenced by different factors related to the reader (e.g. motivation, disposition, and experience), the text (e.g., different text formats or sources), and the tasks or items (e.g., item difficulty). The PISA reading literacy assessment is built on the following subscales:

1. Text formats:
   - *Continuous texts*: typically composed of sentences that are, in turn, organized into paragraphs (newspaper reports, essays, novels, short stories, reviews, and letters).
   - *Noncontinuous texts*: most frequently organized in a matrix format (lists, tables, graphs, diagrams, advertisements, schedules, catalogues, indexes, and forms).
   - *Mixed texts*: combinations of continuous and noncontinuous texts.
2. Text sources (units of texts):
   - *Single-source texts*: have a definite author, a time of writing or publication date, a title or reference number, or are presented to the reader in isolation from other texts (even if there is no source indication). Require students' literal and inference comprehension as well as requiring them to scan and locate, assess the quality and credibility of texts, and reflect on content and form.
   - *Multiple-source texts*: have different authors, are published at different times, or have different titles or reference numbers. Require students' inference comprehension as well as requiring them to search and select relevant text and corroborate or handle conflict.
3. Cognitive processes which are involved in purposeful reading activities and unfold in single or multiple text environments:
   - *Locate information*: related to tasks that require students to search and select relevant texts and access relevant information within texts.
   - *Understand*: related to tasks that require students to represent the explicit meaning of texts as well as integrate information and generate inferences.
   - *Evaluate and reflect*: related to tasks that require the student to assess the quality and credibility of information, reflect on the content and form of a text, and detect and handle conflict within and across texts.

Analysis and modeling approaches

Gender DIF analysis

IRT models. For establishing a reading literacy scale which is comparable across female and male students, we utilized the IRT scaling approach that was introduced in PISA 2015 and used again for PISA 2018. This approach is a multiple-group concurrent cali-

bration with a partial invariance assumption based on the Rasch model (Rasch, 1960) and two-parameter logistic model (2PLM; Birnbaum, 1968) for dichotomous items and the partial credit model (PCM; Masters, 1982) and generalized partial credit model (GPCM; Muraki, 1992) for polytomous items.

PISA needs to establish a comparable scale across different assessment cycles (i.e. over time), administration modes (PBA and CBA), countries, and languages. Therefore, trend items are administered together with newly developed items in each cycle. Moreover, trend items developed before PISA 2015 exist in paper- and computer-based versions (note that no new paper-based items have been developed since 2015; PBAs are for transitioning reasons only). While the trend items allow for linking the current assessment to previous assessment cycles, new items reflect the updated PISA framework for the major domain. Both trend and new item parameters need to be estimated on the same common scale, which is achieved through the IRT scaling. While trend items developed before the 2015 cycle were scaled with a hybrid model that combines all four IRT models, all items developed for PISA 2015 and since were scaled with the 2PLM and GPCM. The hybrid model was introduced to allow a smooth transition from the Rasch model and PCM (which were used for PISA historically before 2015) to the 2PLM and GPCM (which were introduced in PISA 2015). For more details on the introduction of this new scaling model, see OECD (2017a) and von Davier, Yamamoto et al. (2019).

PISA 2018 DIF modeling approach. To account for item-by-country and item-by-language interactions in PISA – that is, DIF resulting from cultural, regional, or language differences – country-by-language groups (i.e., countries divided into different test languages if the sample sizes allow it) are used as a grouping variable in the IRT scaling (multiple-group concurrent calibration). The IRT analysis starts with a full invariance assumption to estimate a common (i.e., comparable) scale. More precisely, only common international item parameters are estimated in this first analysis step, with items receiving the same item parameter estimates across the different groups. Using item-fit statistics, the fit of these common item parameter estimates is evaluated in each country-by-language group. In the case of misfit due to DIF, unique country-specific item parameters are estimated in a stepwise procedure (see more information about this approach in OECD [2017a, 2020a] and in Lee & von Davier [2020] in the first volume of this special issue).

If the magnitude and direction of DIF is the same in a group of countries, the same unique parameter is estimated for all affected countries; hence, we could call them unique group-specific parameter estimates. Eventually, this results in a model with partial invariance where most items receive common item parameter estimates, while a subset receive unique item parameter estimates. Items with unique parameter estimates are removed from the maximum likelihood estimation for estimating common international item parameters but still provide additional information for the ability estimation at the level of countries. The inclusion of unique item parameter estimates for only a subset of items slightly reduces the comparability of the scale, but does not present a threat for cross-country comparisons as the majority of items receive common international item parameter estimates.

With regard to establishing a comparable scale across administration modes for trend items in PISA 2015, when the majority of countries moved to a CBA, the same partial invariance approach was used. A specifically designed mode effect study was conducted in PISA 2015 that identified a subset of items which were easier or harder on computer and required different parameters (OECD, 2017a; von Davier, Khorramdel et al., 2019). In summary, PISA 2015 established a scale which is comparable across different administration modes, assessment cycles, countries, and languages. PISA 2018 utilized the PISA 2015 item parameter estimates for trend items and estimated the new items (developed for PISA 2018) on the same scale. DIF was evaluated for new and trend items and accounted for as described.

Gender DIF modeling approach. In the analysis of the present study, we utilized the official (common and unique) PISA 2018 main survey item parameter estimates in our IRT models for trend and new items to examine gender DIF. Because these item parameter estimates already account for mode-, country-, and language-specific DIF, our analysis focused on gender DIF to establish a scale that allows fair comparisons across gender groups. We used the same IRT scaling approach, that is, a multiple-group IRT model (based on the Rasch model/PCM and 2PLM/GPCM) with partial invariance and conducted a concurrent calibration. However, in our case, we used country-by-language-by-gender groups (232 groups) as the grouping variable in the multiple-group IRT model to evaluate gender DIF. This resulted in smaller sample sizes per group compared to the model used in PISA (with country-by-language groups only). Some groups were too small to compute reliable item-fit statistics (note that we compute item-fit statistics only if more than 250 responses are available for an item). This model allowed us to examine gender DIF after accounting for item-by-country and item-by-language interactions for groups with a large enough sample size.

In the first step, we assumed the same item parameters across gender groups in the IRT model and evaluated the fit of these parameters for each item in each country-by-language-by-gender group using item-fit statistics (details about the fit statistics and the evaluation criteria are presented later). Item misfit was interpreted as gender DIF. In subsequent steps, we estimated unique gender-group specific item parameters in the case of DIF. As in PISA 2015 and 2018, we scored omitted items (i.e., items with no valid responses in the middle of the test session, followed by valid responses to subsequent items) as incorrect responses, while not-reached items (i.e., items with no responses at the end of the test session which are not followed by a valid response) were treated as not administered and not included in the likelihood estimation. Based on the results of our analysis, we updated the PISA 2018 item parameter estimates and, thus, the reading literacy scale. That is, we included additional gender-group specific unique item parameter estimates to account for gender DIF. These updated item parameter estimates were then used in our subsequent analysis for examining gender differences based on a reading literacy scale which is comparable across female and male students.

Diagnostic value of reading literacy subscales

A confirmatory multidimensional IRT (MIRT) framework with fixed item parameters was used to investigate the usefulness of splitting the overall reading literacy test score into subscores that account for text dimensions (text formats, text sources) and cognitive reading processes when examining gender differences. We fit two 3-dimensional IRT models (3D models): one to account for the different text formats (continuous, noncontinuous, and mixed formats) and one to account for different cognitive processes (locate information, understand, evaluate and reflect). A 2-dimensional IRT model (2D model) was fit to account for different text sources (single and multiple sources). The item parameters in the different MIRT models were not estimated freely but fixed to the values obtained from the unidimensional IRT model (1D model) that was used to examine gender DIF. More precisely, the 1D model with updated 2018 item parameter estimates which included additional gender-group specific unique item parameter estimates to account for gender DIF. This approach allows us to obtain subscale scores in a manner that retains the reading literacy scale. Moreover, we followed the PISA approach where a unidimensional scale is assumed for reading literacy and test scores (plausible values)[8] for subscales that are calculated based on the item parameter estimates obtained from a unidimensional IRT model (OECD, 2017a, 2020a). The rationale behind this approach is the assumption that a unidimensional scale describes the international data better than a multidimensional scale since subscales are highly correlated. Hence, item parameter estimates based on a unidimensional scale are assumed to provide reliable cross-country comparisons at the international level. Nevertheless, the produced subscale scores might still be informative at the national or country level.

To examine whether the assumption of unidimensionality holds in the 2018 data, we estimated all described MIRT models (1D, 2D, 3D) without fixing the item parameters. We used the PISA 2018 item parameter estimates as starting values in the estimation but estimated all item parameters freely. We estimated the item parameters to be common across the 116 country-by-language groups and not accounting for any DIF. A comparison of the resulting model fit indices showed a unidimensional reading literacy scale can be assumed at the international level and that our approach of fixing the updated 2018 item parameters in the MIRT models for producing gender scores is justified (model fit indices are presented in the result section).

In the following, the concept of MIRT models is illustrated with the example of the 2PLM.[9] In MIRT models, the 2PLM can be specified for multiple scales. It is assumed

[8] Plausible values are multiple imputations drawn from a posterior distribution obtained from a latent regression model (also referred to as population modeling or conditioning model) using IRT item parameters from the cognitive PISA assessment and principal components from the PISA student background questionnaire. In PISA, each respondent receives 10 plausible values for each cognitive domain that can be used as test scores to produce group-level statistics (never as individual test scores). For more information on plausible values and population modeling see Mislevy and Sheehan (1987), von Davier et al. (2009), von Davier et al. (2006), or OECD (2017, 2020a).

[9] Note that the 2PLM can be reduced to the Rasch model when setting all slope parameters to one, and the GPCM reduces to the 2PLM when applied to dichotomous data.

that the 2PLM holds, with the qualifying condition that it holds with a different person parameter for each of a set of distinguishable subsets (scales) of items (Reckase, 2009; von Davier et al., 2007). For the case of a multidimensional 2PLM with between-item multidimensionality (each item loads on only one scale), the probability of response ($X_{iv}=1$) to item i in scale k by respondent v can be defined as:

$$P\left(x_{iv}|\theta_v,\beta_i,\alpha_i\right)=\frac{\exp\left[\sum_{k=1}^{K}\alpha_{ik}\left(x_{iv}\theta_{vk}-\beta_i\right)\right]}{1+\exp\left[\sum_{k=1}^{K}\alpha_{ik}\left(x_{iv}\theta_{vk}-\beta_i\right)\right]}, \quad (1)$$

where $\boldsymbol{\theta}_v$ is a vector of latent variables and α_{ik} is the item loading for item i on scale k, with the restriction that each item loads on only one scale. Unidimensional IRT models used in our analysis might be treated as a special case of MIRT where $\boldsymbol{\theta}_v = \theta_v$, that is, one latent dimension is assumed (K=1).

If it is additionally assumed that the 2PLM holds for multiple populations simultaneously (multigroup model), that is, the 2PLM holds with a different set of parameters in different populations, $c=1,..,C$ classes are assumed:

$$P\left(x_{iv}|\theta_{vc},\beta_{ic},\alpha_{ic},c\right)=\frac{\exp\left[\sum_{k=1}^{K}\alpha_{ikc}\left(x_{iv}\theta_{vkc}-\beta_{ic}\right)\right]}{1+\exp\left[\sum_{k=1}^{K}\alpha_{ikc}\left(x_{iv}\theta_{vkc}-\beta_{ic}\right)\right]} \quad (2)$$

We compared the reading literacy proficiency of female and male students based on their weighted likelihood estimates (WLEs; Warm, 1989) obtained from the 1D, 2D, and 3D models across all countries as well as the different country-by-language groups. All WLEs were transformed to the PISA scale using the transformation coefficients for reading literacy (OECD, 2017b) reported in PISA 2015 and PISA 2018 (A = 131.5806, B = 437.9583):

$$WLE_T = A \times WLE_U + B \quad (3)$$

The subscripts T and U correspond to the transformed and untransformed values, respectively. Note that plausible values obtained from the population model are used in PISA, not the WLEs obtained from the IRT scaling, for providing the PISA transformed scores.

Attitudes towards reading

To examine the impact of students' attitudes towards reading on reading proficiency and their interaction with gender, we used linear regression models, including selected BQ variables as regressor variables (see Table 1). For each BQ variable, we utilized the WLEs provided in the official PISA 2018 dataset on the OECD website. The proficiency measure for reading literacy (PV1) was obtained from the official PISA 2018 dataset as well.

Table 1:
Variables from the PISA 2018 Student BQ Used in the Regression Model

Gender	
Gender = 0	Female
Gender = 1	Male
Reading attitudes and teacher stimulation	
Stimread	Teacher's stimulation of reading engagement perceived by student (WLE)
Joyread	Joy of reading (WLE)
Screadcomp	Self-concept of reading: Perception of competence (WLE)
Screaddiff	Self-concept of reading: Perception of difficulty (WLE)

The regression model was applied to the PISA 2018 data for selected countries with either large or small gender differences in the unidimensional reading literacy scale. We ran the regression model for multiple country-by-language groups, but selected only a few for presentation purposes. We selected Germany, Korea, and the United States as groups with smaller gender differences and Finland, Israel-Arabic, and Israel-Hebrew as groups with larger gender differences. For each group, we fit the following regression model:

$$\begin{aligned} Reading\ PV1 = {} & \beta_0 + \beta_1 Gender + \beta_2 Stimread + \beta_3 Joyread \\ & + \beta_4 Screadcomp + \beta_5 Screaddiff + \beta_6 Gender * Stimread \\ & + \beta_7 Gender * Joyread + \beta_8 Gender * Screadcomp \\ & + \beta_9 Gender * Screaddiff + e_v \end{aligned} \tag{4}$$

where PV1 denotes the reading proficiency measured with the first plausible value as the dependent variable, β_0 and β_1 to β_9 denote the regression parameters (the intercept and slopes respectively), and e_v describes the independent error term.

Item and model fit indices in the IRT analysis

Following the approach taken in PISA, we used the root mean square deviation (RMSD)[10] and the mean deviation (MD) as item-fit statistics for examining item parameter invariance and as reference indices for allowing unique item parameter estimates in instances of gender DIF. Both fit statistics quantify the magnitude and direction of deviations in the observed data from the estimated item characteristic curves (ICC) and provide complementary information. While the MD is more sensitive to deviations of ob-

[10] We thank an anonymous reviewer for pointing out that the RMSD might potentially have issues in detecting DIF in items for which most respondents in a country have a very low (or high) probability of providing a correct answer (Tijmstra, et al.,2019). However, this might not be an issue for the PISA 2018 reading literacy data as the assessment is based on a MSAT design.

served item difficulty parameters from the estimated ICC, the RMSD is sensitive to the deviations of both the item difficulty parameters and item slope parameters. In contrast to other measures for the evaluation of model-data fit, such as INFIT and OUTFIT measures under the Rasch model, the MD and RMSD indices are not affected by sample size and are available for a range of IRT models. See, for example, Khorramdel et al. (2019) or OECD (2017a) for more details on the MD and RMSD.

In the IRT scaling, MD and RMSD are computed for each item in each group. Choosing a specific item-fit threshold can be rather subjective and varies between studies. While, usually, MD values ≥ .20 and ≤ -.20 (values close to zero indicate perfect item fit) and RMSD ≥ .20 are considered for examining item misfit in most studies, some studies based on ILSA data (Oliveri & von Davier, 2011, 2014) used an even stricter criterion (MD values ≥ .10 and ≤ -.10 and RMSD ≥ .10), and in the PISA 2015 and 2018 main survey scaling, a threshold of MD values ≥ .12 and ≤ -.12 and RMSD ≥ .12 was used (OECD, 2017a, 2020a).

In our gender DIF analyses, we followed the operational procedures in PISA and used a threshold of RMSD ≥ .12 for estimating unique item parameters. Hence, items that showed misfit to the common item parameter estimates based on these thresholds received a new unique (gender-specific) item parameter estimate. It is important to note that in the IRT models of the current study, MD and RMSD values were not estimated for item-by-group combinations if the response rate in a certain group was < 250 (for a particular item). This was done to ensure that decisions about the estimation of unique item parameters were not biased by small sample sizes.

For overall model-data fit evaluation, we used the Akaike Information Criterion (AIC; Akaike, 1974) and the Bayesian Information Criterion (BIC; Schwarz, 1978). While the use of the AIC is recommended for conditions with small sample sizes, where errors occur from underfitting a model, the BIC is more appropriate in conditions with large sample sizes, where errors occur from overfitting a model (Dziak et al., 2012). Note that smaller AIC and BIC values indicate better fit.

Software for the IRT analysis

All IRT analyses were run using the software *mdltm* (von Davier, 2005[11]), which was also used in the operational PISA 2015 and 2018 analyses. The *mdltm* software allows the application of the mixture general diagnostic modeling framework (MGDM), which includes multigroup IRT models based on the Rasch model/PCM and the 2PL/GPCM as special cases (von Davier, 2010). The software provides marginal maximum likelihood estimates obtained using customary expectation-maximization methods. It was designed to handle large datasets as well as complex test and sampling designs. It allows for the estimation of a number of different latent variable models, includes different constraints for parameter estimation, and provides different model and item-fit statistics as well as methods for proficiency estimation. In addition, it can handle missing data by design and

[11] Note that we used a software update from 2019.

nonresponse, as well as multiple populations and weights to account for complex sampling (for a detailed description of the software, see, for example, Khorramdel et al., 2019).

Results

Dimensionality of the PISA 2018 reading literacy scale

The results of the 1D and MIRT models with 116 country-by-language groups and freely estimated common item parameters – without accounting for DIF – are presented in Table 2. For the MIRT models, we estimated a 2D model to account for the different text sources (2D-Source), a 3D model to account for the different text formats (3D-Format), and a 3D model to account for the different cognitive processes (3D-Cognitive). Results based on the AIC and BIC show a slightly better model-data fit for the MIRT models compared to the 1D model. However, the difference in model fit improvement based on the Gilula & Haberman (1994) log penalty measure (the negative expected log likelihood per observation) is negligible. The more restrictive 1D model reaches 98.70% of the likelihood improvement compared to the more general 3D-Format model, both in reference to improvement over the independence (baseline) model. Moreover, the model-based correlations between the subscales in each MIRT model are high across the different country-by-language groups, suggesting there is a single identifiable underlying latent variable. The model-based correlations (correlations of skill distributions) ranged from 0.75-0.93 in the 2D-Source model, from 0.85-0.99 in the 3D-Cognitive model, and from 0.81-0.96 in the 3D-Format model (only two of the groups show medium correlations in the 3D models as low as 0.42-0.69). Moreover, the parameter estimates for all models still include (potential) bias from country, language, and gender DIF since no unique item parameters were estimated. DIF usually can be a potential source of unintended multidimensionality. Therefore, we are confident that the 1D model should be preferred and that using item parameter estimates from a unidimensional reading literacy scale for examining gender DIF and producing subscale scores to investigate gender differences is a reasonable approach.

Gender DIF and comparability of the reading literacy scale

The PISA 2018 reading literacy scale was established as a unidimensional scale comparable across different countries, languages, administration modes, and assessment cycles. According to the PISA 2018 technical report, the majority of the item parameter estimates on which the scale is based consists of common international parameter estimates (87.70% for trend items, 88.39% for new items), which are invariant across countries, and a subset of items received country- or language-specific item parameter estimates (12.30% for trend items, 11.40% for new items), which are noninvariant across countries. Hence, the final scale already accounts for item-by-country and item-by-language interactions DIF (for detailed results, see OECD [2020b]).

Table 2:
Overall Model Fit for Unidimensional and Multidimensional IRT Models with 116 Country-by-Language Groups and Freely Estimated Item Parameters

Model	Likelihood	Deviance	AIC	BIC	Model-based log penalty per item	% Improvement
Independence					0.6493 0	0.00
1D	-7411772.7 6	14823545.5 2	14825677.5 2	14837275.9 2	0.5633 2	98.70
2D Source	-7404413.8 7	14808827.7 3	14811883.7 3	14828508.8 4	0.5627 6	99.34
3D Cognitive	-7401046.7 8	14802093.5 6	14806305.5 6	14829219.4 9	0.5625 1	99.64
3D Format	-7396898.5 3	14793797.0 5	14798009.0 5	14820922.9 8	0.5621 9	100.00

Note: The item parameters in all models are common parameters across country-by-language groups and were freely estimated; the official PISA 2018 common and unique parameters were used as starting values in the estimation.

Table 3:
Percent of Item-by-Group Interactions (Gender DIF) and Comparability of Item Parameter Estimates Obtained from the 1D model with 232 Country-by-Language-by-Gender Groups and Fixed Item Parameters

1D Model	Gender DIF (RMSD ≥ 0.12)			Comparability		
	All items	MAC items	HUM items	All items	MAC items	HUM items
Total	0.83%	0.48%	1.64%	99.17%	99.52%	98.36%
Female	0.71%	0.47%	1.28%	99.29%	99.53%	98.72%
Male	0.93%	0.49%	1.95%	99.07%	99.51%	98.05%

Note: MAC refers to machine-coded items (such as multiple-choice items and short text responses) and HUM refers to human-coded items (longer text responses). The presented percentages in the table refer to the number of estimated item parameters with RMSD ≥ 0.12 in relation to item parameter estimates with RMSD below that threshold, not to the number of items with DIF.

To examine gender DIF, we estimated a 1D model with reading literacy as a unidimensional scale (i.e., all items assigned to the same scale) based on the current PISA IRT scaling approach. The 1D model used country-by-language-by-gender as a grouping variable, resulting in 232 groups, and we fixed the item parameters to the official PISA 2018 (common and unique) item parameter estimates for both female and male students (i.e., we assumed invariance of item parameters across gender groups). Then, we evaluated DIF for each item in each group based on the RMSD (values ≥ 0.12 were considered as DIF). The DIF results are illustrated in Table 3 and figures 1 and 2.

Results presented in Table 3 show the percentage of item-by-group interactions for the combination of 346 items x 232 gender groups (percentage of gender DIF), and the resulting comparability of the reading literacy scale across gender groups defined as 100 minus the percentage of gender DIF.[12] For example, if the amount of gender DIF is 0.83% across all items, the comparability of the scale based on all items is 99.17%. Results showed only a small percentage of gender DIF overall, with a tendency of open-ended response items coded by human raters (HUM) to show a higher percentage of DIF than machine-coded items (MAC; multiple-choice items and short text responses). There was a slightly higher percentage of DIF in HUM items for male than female students. However, differences seemed too small to be meaningful, and the overall comparability of item parameter estimates was still high, with >99% for MAC items and >98% for HUM items. Interestingly, the country-by-language groups with the highest average gender differences in reading proficiency (WLEs) were not necessarily the groups with the highest number of items showing gender DIF; we could not find any meaningful pattern in this regard.

In a second step, we estimated unique gender-specific item parameters in case of gender DIF in the 1D model. Therewith, we updated the PISA 2018 item parameter estimates by including additional unique gender group-specific parameter estimates. The overall model-fit statistics presented in Table 4 indicated a better model-data fit (like expected) for the model that now accounted for gender DIF (1D*), compared to the model which did not account for gender DIF (1D). (Also note that the 1D model and 1D* model with fixed item parameters clearly show a better model data fit than the 1D model with freely estimated item parameters in Table 2.) Comparing our findings in Table 3 to the PISA 2018 findings showed that the percentage of country- and language-specific DIF was higher than gender-specific DIF. Moreover, the correlation between WLEs obtained from the 1D model before and after accounting for gender DIF (i.e., the same model without and with gender-specific unique item parameter estimates) was $r = 0.99$. Thus, it might not make a real difference to account for gender DIF in our empirical example. However, for examining gender differences in the next section of the paper with multidimensional IRT models, we used the updated item parameter estimates obtained from the 1D* model.

[12] A detailed table which illustrates the item-by-group interactions (i.e. which items show an RMSD ≥ 0.12 in which country-by-language-by-gender group) in the 1D model, including specific item IDs, is not provided in this paper but can be requested from the authors.

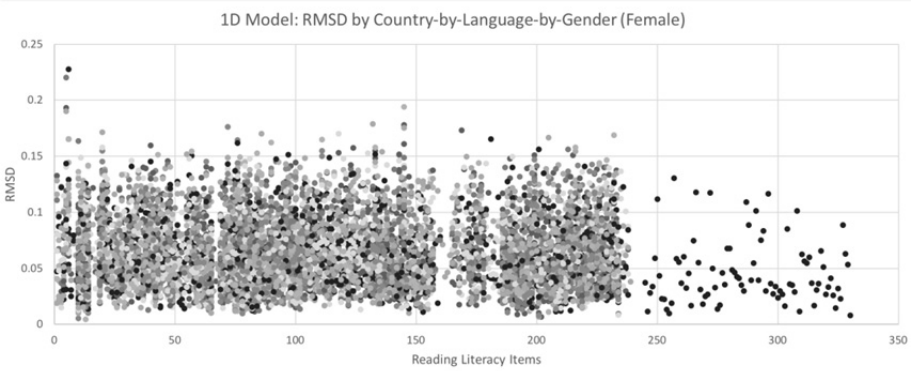

Figure 1:
RMSD values for all items in each country-by-language-by-gender group (Female) in the 1D model; RMSD values for trend items in PBA countries are presented at the very right (note that PBA countries did include trend items only); sparse areas within CBA countries are related to missing RMSD values due to too few responses (note that RMSD values were calculated for items with ≥ 250 responses in a particular group only).

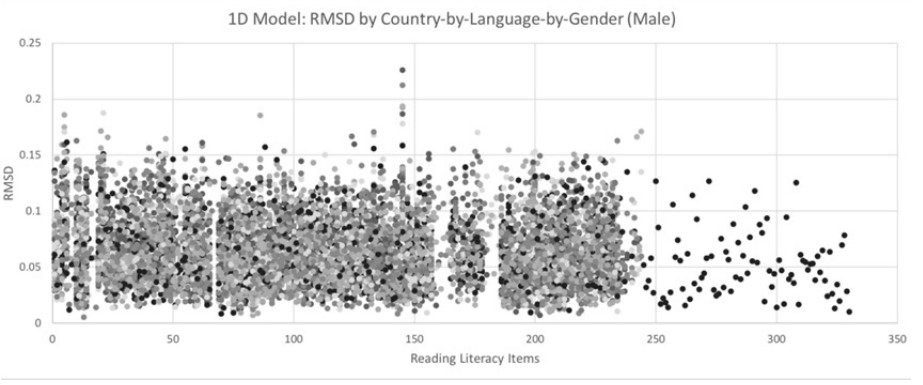

Figure 2:
RMSD values for all items in each country-by-language-by-gender group (Male) in the 1D model; RMSD values for trend items in PBA countries are presented at the very right (note that PBA countries included trend items only); sparse areas within CBA countries are related to missing RMSD values due to too few responses (note that RMSD values were calculated for items with ≥ 250 responses in a particular group only).

Table 4:
Overall Model Fit for Unidimensional and Multidimensional IRT Models with 232 Country-by-Language-by-Gender Groups and Fixed Item Parameters

Model	Likelihood	Deviance	AIC	BIC	Model-based log penalty per item	Independence log penalty per item
1D	-7262954.1 4	14525908.2 9	14527762.2 9	14537848.3 3	0.5520 1	0.6493 0
1D*	-7249050.7 3	14498101.4 7	14501657.4 7	14521002.6 5	0.5509 6	0.6493 0
2D-Source	-7281638.27	14563276.54	14566978.54	14587117.99	0.55343	0.64930
3D-Format	-7279453.0 9	14558906.1 8	14564924.1 8	14597663.0 2	0.5532 7	0.6493 0
3D-Cognitive	-7282893.4 5	14565786.8 9	14571816.8 9	14604621.0 2	0.5535 3	0.6493 0

Note: The 1D model does not account for gender DIF; the 1D* model does account for gender DIF by including new gender-specific unique item parameter estimates. (Also note that the reported model fit indices should not be interpreted in terms of dimensionality as the MIRT models are based on the 1D* model item parameter estimates and, therefore, cannot show a better model fit than the 1D* model. For model comparisons, regarding the dimensionality of the scale, see Table 2.)

Diagnostic value of the reading literacy subscales

For examining gender differences in the reading literacy subscales, we fit MIRT models with 232 country-by-language-by-gender groups with all item parameters fixed to the estimates obtained from the 1D* model, which accounts for gender DIF. Again, we fit a 2D model to account for the different text sources (2D-Source), a 3D model to account for the different text formats (3D-Format), and a 3D model to account for the different cognitive processes (3D-Cognitive). All resulting WLEs (person parameters) were transformed onto the PISA scale using the PISA 2015 transformation coefficients (OECD, 2017b).

Results are illustrated for different levels. First, we present results at the aggregated level, that is, over all countries. Then, we discuss results in more detail at the country-by-language level. The gender differences at the subscale level (MIRT models) were compared to the gender differences in the main reading literacy scale (1D* model) to explore whether the subscale level provided more information than the main reading scale alone. Results at the aggregated level showed a similar pattern of gender differences in all subscales and the main scale. That is, female students outperformed male students (see the group means based on WLE_T values in Table 5). Moreover, the effect sizes of the gender differences for each scale based on Cohen's *d* were very similar as well, with values close to $d \approx 0.30$ (see Table 5).

Results at the country-by-language level showed a similar picture. We found similar gender differences at the subscale level, relative to the main reading literacy scale. That is, female students outperformed male students in all country-by-language groups (see Tables 7–10 in the supplemental material). Moreover, correlations of the group-level gender differences between the main reading scale and the different subscales (based on WLEs) were high (> 0.9).[13] This indicated that groups with high or low gender differences in the unidimensional model showed low or high gender differences in the single subscales as well, respectively.

If we look at the group-level effect sizes for the gender differences, we see very similar patterns between results for the main reading literacy scale and results for each subscale. For all scales, the same countries showed small or large effect sizes for the gender differences, respectively. In general, we saw the largest gender differences and effect sizes in groups with Arabic language (for detailed results, see Tables 7–10 of the supplemental material). We could also see similar patterns across the models with regard to the correlation of the average WLEs by country with the mean gender difference by country. A negative low correlation was observed for the main reading literacy scale and each subscale, indicating larger gender differences for countries with lower mean reading profi-

[13] Correlations between mean gender differences (based on WLEs) from the reading literacy scale (1D* model) and the subscales are as follows: single-text sources ($r = 0.98$), multiple-text sources (2D-Source model; $r = 0.97$), continuous text formats ($r = 0.98$), noncontinuous test formats ($r = 0.90$), mixed text formats (3D-Format model; $r = 0.92$), locate information ($r = 0.92$), understand ($r = 0.97$), evaluate and reflect (3D-Cognitive model; $r = 0.93$).

Table 5:
Effect Sizes (Cohen's d) of Mean Gender Differences Obtained from the IRT Models with 232 Country-by-Language-by-gender Groups Averaged across all Groups

Model	Scale/Subscale	Group	Mean	SD	N	Cohen's d
1D*	Reading Literacy	Female	470.54	87.68	194872.5 0	0.34
		Male	439.29	97.43	197633.90	
2D-Source	Single text sources	Female	470.11	93.74	194872.5 0	0.34
		Male	435.92	105.50	197633.93	
	Multiple text sources	Female	472.18	93.70	194872.5 0	0.29
		Male	443.34	104.89	197633.93	
3D-Format	Continu ous text formats	Female	472.48	94.53	194872.50	0.34
		Male	438.90	105.52	197633.93	
	Noncontinuous text formats	Female	468.47	97.07	194872.5 0	0.28
		Male	439.04	109.80	197633.93	
	Mixed text formats	Female	469.43	95.90	194872.5 0	0.27
		Male	441.70	108.52	197633.93	
3D-Cognitive	Locate information	Female	470.51	94.28	194872.5 0	0.32
		Male	438.19	107.86	197633.93	
	Understand	Female	471.12	93.56	194872.50	0.32
		Male	439.02	105.85	197633.93	
	Evaluate and reflect	Female	471.62	96.34	194872.5 0	0.29
		Male	442.06	105.03	197633.93	

ciencies (1D* model: r = -0.19; 2D-Source model: r = -0.19, r = -0.23; 3D-Format model: r = -0.22, r = -0.20, r = -0.19; 3D-Cognitive model: r = -0.20, r = -0.23, r = -0.16). However, the correlations were not large, and we observed higher-performing countries showing higher gender differences than lower-performing countries in some cases. The main observation here is that we do not observe different patterns of correlations for subscales compared to the main reading literacy scale.

Attitudes towards reading

The results of the regression analysis (see Table 6) showed a significant effect (p-value < 0.05) for gender in all selected countries, which was expected given the observed gender differences. This indicated that gender was likely to be a meaningful addition to the regression model for explaining reading proficiency. With regard to the reading attitude-related variables and the interaction between these variables and the gender variable, the results differed between countries.

In the United States (USA), most reading attitudes were not significant predictors and there were no significant interactions between gender and reading attitudes. In Korea (KOR), all reading-attitude variables showed significant effects, but no significant interaction effects were observed between gender and reading attitudes. In Germany (DEU), most reading-attitude variables showed significant effects; there were significant interaction effects between gender and the variables "Stimread" (perceived teacher's stimulation of reading engagement) and "Joyread" (joy of reading).

In Israel (ISR), almost no attitude variables and none of the interactions showed significant effects in the Arabic language (similar to results in the Arabic-speaking countries Qatar and United Arab Emirates, both of which are not presented in this paper),[14] while we did see significant effects for two attitude variables and most of the interactions between attitude variables and gender in Hebrew. In Finland (FIN), almost all attitude variables were significant predictors; there was also one significant interaction effect between gender and the variable "Joyread" (joy of reading).

These results illustrate that attitudes towards reading can partly explain reading proficiency scores in some countries, while other countries showed no significant results. The interaction effects indicate that, at least in some countries, gender differences in reading proficiency could partly be explained by gender differences in reading attitudes. However, results varied from country to country and cannot be generalized across countries. We also saw no meaningful patterns between countries with smaller gender differences in reading proficiency compared to countries with larger gender differences.

[14] Result tables can be requested from the authors.

Table 6:
Regression Analysis Results for the United States (USA), Korea (KOR), Germany (DEU), Israel (ISR), and Finland (FIN)

	USA-English				**KOR- Korean**				**DEU-German**			
	Coef	*Std.Err*	*t val*	*Pr(>|t|)*	*Coef*	*Std.Err*	*t val*	*Pr(>|t|)*	*Coef*	*Std.Err*	*t val*	*Pr(>|t|)*
(Intercept)	515.41	2.22	231.74	0.00	527.07	1.80	293.20	0.00	527.20	2.19	240.82	0.00
Gender	-20.92	3.14	-6.67	0.00	-19.39	2.50	-7.77	0.00	-21.83	3.02	-7.23	0.00
Stimread	-0.39	0.22	-1.77	0.08	-1.04	0.41	-2.56	0.01	-1.01	0.12	-8.29	0.00
Joyread	0.31	0.26	1.20	0.23	1.57	0.50	3.14	0.00	1.10	0.13	8.21	0.00
Screadcomp	0.05	0.17	0.32	0.75	0.46	0.20	2.33	0.02	-0.05	0.14	-0.35	0.73
Screaddiff	-0.90	0.18	-5.08	0.00	-1.28	0.23	-5.53	0.00	-0.69	0.13	-5.39	0.00
Gender* Stimread	0.48	0.28	1.69	0.09	-0.11	0.55	-0.21	0.84	0.54	0.16	3.43	0.00
Gender* Joyread	-0.30	0.32	-0.95	0.34	-0.37	0.64	-0.58	0.56	-0.45	0.17	-2.67	0.01
Gender* Screadcomp	-0.23	0.22	-1.03	0.30	-0.27	0.26	-1.03	0.30	-0.27	0.20	-1.36	0.18
Gender* Screaddiff	-0.04	0.23	-0.17	0.86	-0.13	0.31	-0.41	0.69	0.10	0.18	0.55	0.58
	ISR-Arabic				**ISR-Hebrew**				**FIN-Finish-Swedish**			
	Coef	*Std.Err*	*t val*	*Pr(>|t|)*	*Coef*	*Std.Err*	*t val*	*Pr(>|t|)*	*Coef*	*Std.Err*	*t val*	*Pr(>|t|)*
(Intercept)	408.61	3.11	131.20	0.00	527.39	2.54	207.82	0.00	550.32	1.80	306.34	0.00
Gender	-44.32	4.75	-9.33	0.00	-29.91	3.60	-8.30	0.00	-50.08	2.54	-19.72	0.00
Stimread	-0.10	0.21	-0.46	0.65	0.13	0.06	1.96	0.05	-1.09	0.16	-6.72	0.00
Joyread	-0.36	0.21	-1.71	0.09	-0.05	0.24	-0.21	0.83	1.14	0.22	5.21	0.00
Screadcomp	-0.12	0.14	-0.85	0.40	-0.68	0.21	-3.32	0.00	-0.01	0.17	-0.08	0.94
Screaddiff	-0.62	0.14	-4.58	0.00	-0.93	0.19	-4.98	0.00	-0.81	0.16	-4.96	0.00
Gender* Stimread	-0.39	0.25	-1.53	0.13	0.23	0.09	2.47	0.01	0.36	0.19	1.85	0.06
Gender* Joyread	0.21	0.26	0.80	0.43	-0.65	0.28	-2.35	0.02	-1.15	0.25	-4.56	0.00
Gender* Screadcomp	-0.19	0.20	-0.97	0.33	0.55	0.26	2.13	0.03	0.34	0.23	1.51	0.13
Gender* Screaddiff	0.36	0.20	1.80	0.07	-0.06	0.24	-0.25	0.80	-0.05	0.23	-0.22	0.83

Discussion

In this paper, we examined gender DIF, differences in reading literacy subscales, and differences in students' attitudes towards reading as potential sources for observed gender differences in the PISA 2018 reading literacy scores across countries and languages. We illustrated the use of a multiple-group concurrent calibration based on the 2PLM and GPCM with a partial invariance approach for examining gender DIF and establishing a comparable reading literacy scale across gender groups. We used the same IRT scaling approach introduced in PISA 2015 and used again in PISA 2018 for placing trend and new items on a common scale in each cognitive domain. This approach allows for establishing comparable scales across countries, languages, administration modes (computer- and paper-based tests) and assessment cycles for reporting trend measures over time. In our study, we extended this scaling approach for examining and treating gender DIF. The utilized scaling approach assumed perfect item parameter invariance across different groups of interest (e.g., country-by-language groups) as a first step in our analysis to estimate common or international item parameters across all groups. Next, item fit was evaluated (based on RMSD and MD indices) for each item in each group to examine DIF, which was defined as misfit to the common item parameter estimates. More precisely, the difference between the empirical ICC and the model-based ICC was evaluated for each item in each group. In instances of misfit, unique country-specific item parameters were estimated and the misfit was removed from the likelihood estimation for estimating common international item parameters. If only a small subset of items received country-specific item parameter estimates in the scaling model, with the majority retaining the common parameters (as is usually the case in PISA), the comparability of countries was decreased only slightly and meaningful cross-country comparisons were still possible. This modeling approach, based on common and unique item parameter estimates, is called the partial invariance approach. In addition to comparable item parameter estimates across countries, languages and assessment cycles, the scaling in PISA 2015 accounted for mode effects (i.e., DIF due to the move from a PBA to a CBA of the majority of countries), and the scaling in PISA 2018 established a link between the new multistage adaptive test (MSAT) design for reading literacy and the past data, which are based on a nonadaptive design. In the gender DIF analysis in our study, we utilized the official PISA 2018 item parameter estimates from the PISA 2018 population model (used to compute plausible values) and evaluated gender DIF by splitting the country-by-language groups additionally by gender.

We found only a small amount of gender DIF (i.e., item-by-gender interactions) and estimated gender group-specific unique item parameters to account for it. In accordance with previous studies (Beller & Gafni, 2000; Lafontaine & Monseur, 2009; Schwabe et al., 2015), open-ended response items coded by human raters, which tended to be more difficult and cognitively demanding (Kubinger et al, 2010), showed a slightly higher percentage of gender DIF than machine-coded items (multiple choice items and short text responses). With regard to the relation between the average size of gender differences and the number of items with DIF, no meaningful patterns could be observed. Moreover, the correlation between person parameters (WLEs) from the 1D model with-

out (1D) and with (1D*) gender-specific item parameter estimates to account for gender DIF was very large (r = 0.99). Altogether, the amount of gender DIF was too small to account for the observed gender differences in the reading literacy scale.

Comparing our gender DIF findings to country and language DIF results from the PISA 2018 operational analysis illustrated in the technical report (OECD, 2017b) showed that the percentage of country- and language-specific DIF was higher than the amount of gender-specific DIF found in our study. Hence, accounting for country and language DIF before examining gender DIF seems to be a reasonable approach. One could argue for a model that accounts for all types of DIF at the same time, that is, a model which uses country-by-language-by-gender groups from the beginning. However, this would only be feasible if the sample sizes for all subgroups were large enough. In our analysis, some groups had sample sizes too small to allow an accurate estimation of item-fit statistics and unique item parameters. This could lead to country- and language-specific DIF being undetected and to a less accurate scale.

By estimating gender-specific unique item parameters to account for DIF, we established a more accurate and gender fair reading literacy scale. Based on this scale, we examined the diagnostic value of three reading literacy subscales (text sources, text formats, and cognitive processes) with regard to gender differences. That is, we investigated whether scores at the subscale level provide additional information about gender differences which could be observed in the main reading literacy scale across the different PISA cycles (with female students outperforming male students in all countries). We used three different MIRT models (2D-Source, 3D-Format, 3D-Cognitive), with all item parameters fixed to the values obtained from the 1D* model, mimicking the approach taken in the PISA population model (where item parameter estimates from the unidimensional IRT model were used to compute plausible values at the subscale level). The person parameters (WLEs) obtained from these models were transformed to the PISA scale using the transformation coefficients provided in the official PISA technical report. We then compared the gender-specific mean WLEs from the MIRT models with the ones from the unidimensional model. Results showed similar patterns of gender differences across all models: female students outperformed male students in all country-by-language groups (especially in countries with Arabic as test language), and the correlations of the group-level gender differences between the main reading scale and the different subscales were high ($r > 0.9$).

We also saw very similar patterns between results for the main reading literacy scale and results at the subscale level with regard to effect sizes (Cohen's d) and the correlation between average WLEs by country and mean gender difference by country (low negative correlations close to $r \approx -0.20$). On average, lower-performing countries tended to show larger gender differences than higher-performing countries. But the correlations were not very strong, and there were higher-performing countries such as Finland with larger gender differences. It is possible that the correlation was influenced by other mediator variables which were not accounted for in the analyses. In summary, the subscale level did not reveal different results nor additional information compared to the main reading literacy scale, in contrast to findings from previous studies on gender differences in

different text formats (OECD & Statistics Canada, 2005; Solheim & Lundetræ, 2018) and cognitive processes required by the reading task (Solheim & Lundetræ, 2018).

The different findings in our study, including those from examining the dimensionality of the reading literacy scale with MIRT models (based on freely estimated item parameters), support the assumption of a unidimensional reading literacy scale in PISA and indicate that the subscales are measuring similar constructs at the international, and likely even at the country level (the latter needs to be further examined with country-level dimensionality analysis). This is reassuring with regard to the use of a single main reading literacy score for cross-country comparisons.

Finally, we investigated the extent to which gender differences in the main reading literacy scale are related to students' attitude towards reading (perceived teacher's stimulation of reading engagement, joy of reading, perception of reading competence, perception of difficulty), as measured with the PISA 2018 student BQ. The results of linear regression analyses for selected countries with either smaller or larger mean gender differences varied across countries. In some countries, we saw significant interaction effects between attitude-related variables and gender, while other countries showed no significant interaction effects. A significant interaction effect between these variables did not seem related to a country having smaller or larger gender differences or to attitude variables showing a significant effect on their own. Hence, in some countries, gender differences in reading proficiency could be partly explained by gender differences in reading attitudes. However, results could not be generalized across countries. There might have been some relation between the test language and finding interaction effects, but more detailed analysis would be needed to verify what, at this point, is merely a working hypothesis.

In summary, the small amount of gender DIF could not explain the observed gender differences in the PISA 2018 reading literacy scale, and neither could scores at the subscale level, which did not seem to provide diagnostic value beyond the main reading literacy score. We found that, at least in some countries, the gender differences might partly be explained by students' attitudes towards reading, but these findings seemed country- and language-specific and could not be generalized across country-by-language groups.

To our knowledge, the presented study is the first comprehensive study to analyze gender DIF in the PISA reading literacy scale on an international level. Up to this point, the operational PISA scaling did not include an evaluation of gender DIF at the international level; rather the assumption was that country- and language-based DIF presented a bigger problem. Our study has now tested and confirmed this assumption for the 2018 data.

Limitations

We acknowledge that fixing the item parameters in the MIRT models to those obtained in a unidimensional model for producing subscale level scores might raise concerns for some. Our rationale behind this approach in PISA was the assumption that a unidimensional scale described the international data better than a multidimensional scale, as subscales were highly correlated. We were able to confirm this assumption based on the

2018 PISA data in our study. Hence, item parameter estimates based on a unidimensional scale were assumed to provide reliable cross-country comparisons at the international level. Moreover, we used this approach only to obtain gender-specific group means at the subscale level and we are not comparing the model fit indices of the models with fixed parameters (which we stress in our note below Table 4). Since reading literacy dimensionality was not examined at the national level, our initial assumption was that the subscale scores produced might still be informative at the national or country level, for example, when examining gender differences. However, the reason we did not find any impact of the subscale level on the observed gender differences might actually be because the PISA reading literacy scale was "too" unidimensional, possibly even at the country level (note that this is just a hypothesis and needs to be examined in a separate study). Similar gender studies could focus on reading scales where multidimensionality holds, and with freely estimated item parameters in MIRT models. In such a case, country, language, and gender DIF would need to be examined for each MIRT model separately. Another possible concern with our approach of fixing the item parameters from a unidimensional model in MIRT models for obtaining group means at the subscale level is the fact that the PISA 2018 reading literacy scale is based on an MSAT design. Jewsbury and van Rijn (2019) described the problem of an incomplete routing information in a unidimensional model as potential source of bias in item parameter estimation. However, the PISA 2018 technical report stresses the comparability of the reading literacy data between the linear and MSAT test design with regard to the percent of correct responses (calculated by standardizing the proportions of adaptive paths), response time, omitted responses, and item cluster position effects. PISA 2018 also uses a rather cautious MSAT design which includes large intact sets of item units (up to eight items per unit) and which controls for possible item position effects, leading to similar item characteristic curves (ICCs) between the most difficult and the easiest testlets or modules (a set of several item units that, when combined across all adaptive stages, constitute the administered assessment) – in other words, the item difficulty level gap between modules is not large – and which was shown to perform well in terms of item parameter recovery (Yamamoto, Shin and Khorramdel, 2018; Yamamoto, Shin and Khorramdel, 2019) enabling the estimation of a unidimensional scale. Nevertheless, this possible issue should be examined further, especially if other data sets are used based on different MSAT designs.

Another limitation of our study was that we performed the regression analysis for selected countries only. Other studies could extend them to all countries and languages in PISA to obtain a more accurate picture. Our results based on BQ variables should also be interpreted with caution as a response-style bias (Khorramdel et al, 2017) might exist in the data; we did not test and account for such a bias. Additional variables as measures for test-taking effort such as response time and omitted response rates might be useful. What is more, our findings are limited to the PISA 2018 reading literacy scale. Similar studies using other large-scale assessments and reading scales would be interesting.

The final remark that needs to be made is the relative nature of DIF (Holland & Wainer, 1993). The proposed analysis assumes that most of the items are free of DIF and that DIF effects can be identified. Although the majority of studies analyzing DIF are based on

this assumption, it is untestable from a statistical point of view. In the unlikely but possible situation where all items show the same direction of DIF, our procedure (and the majority of DIF procedures) would not be able to detect it because all DIF will be accounted for by differences in abilities. There are some methods that are trying to cope with this (Bechger & Maris, 2015) but they make different assumptions about the nature of DIF.

Acknowledgements

The work of the second author has been prepared under the project Scales Comparability in Large-Scale Cross-Country Surveys, which is funded by the Polish National Science Centre, as part of the grant competition Sonata 8 (UMO-2014/15/D/HS6/04934).

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723. http://doi.org/10.1109/TAC.1974.1100705

Baker, D. P., & Jones, D. P. (1993). Creating gender equality: Cross-national gender stratification and mathematical performance. *Sociology of Education*, *66*(2), 91–103. https://doi.org/10.2307/2112795

Bechger, T. M., and Maris, G. (2015). A Statistical Test for Differential Item Pair Functioning. *Psychometrika* 80 (2), 317–40. https://doi.org/10.1007/s11336-014-9408-y

Beller, M., & Gafni, N. (1996). 1991 International Assessment of Educational Progress in Mathematics and Sciences: The gender differences perspective. *Journal of Educational Psychology*, *88*(2), 365. https://doi.org/10.1037/0022-0663.88.2.365

Beller, M., & Gafni, N. (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles*, *42*(1–2), 1–21. http://doi.org/10.1023/A:1007051109754

Birnbaum, A. (1968). Some latent trait models and their use in inferring a student's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Borgonovi, F. (2016). Video gaming and gender differences in digital and printed reading performance among 15-year-olds students in 26 countries. *Journal of adolescence*, *48*, 45-61. https://doi.org/10.1016/j.adolescence.2016.01.004

Borgonovi, F., Choi, Á., & Paccagnella, M. (2018). *The evolution of gender gaps in numeracy and literacy between childhood and adulthood*, OECD Education Working Paper, 27. https://doi.org/10.1787/0ff7ae72-en

Borgonovi, F., & Greiff, S. (2020). Societal level gender inequalities amplify gender gaps in problem solving more than in academic disciplines. *Intelligence*, https://doi.org/10.1016/j.intell.2019.101422

Bornstein, M. H., Hahn, C.-S., & Haynes, O. M. (2004). Specific and general language performance across early childhood: Stability and gender considerations. *First Language*, *24*(3), 267–304. https://doi.org/10.1177/0142723704045681

Breda, T., & Napp, C. (2019). Girls' comparative advantage in reading can largely explain the gender gap in math-related fields. *Proceedings of the National Academy of Sciences*, *116*(31), 15435-15440. https://doi.org/10.1073/pnas.1905779116

Buchmann, C., DiPrete, T. A., & McDaniel, A. (2008). Gender inequalities in education. *Annual Review of Sociology*, *34*, 319–337. https://doi.org/10.1146/annurev.soc.34.040507.134719

Camarata, S., & Woodcock, R. (2006). Sex differences in processing speed: Developmental effects in males and females. *Intelligence*, *34*(3), 231–252. https://doi.org/10.1016/j.intell.2005.12.001

Cohen, J. (1977). *Statistical power analysis for behavioral sciences* (revised ed.). Academic Press. http://www.utstat.toronto.edu/~brunner/oldclass/378f16/readings/CohenPower.pdf

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment*, *8*, 69–82. https://eric.ed.gov/?id=EJ1062839

Dziak, J. J., Coffman, D. L., Lanza, S. T., & Li, R. (2012). *Sensitivity and specificity of information criteria*. Methodology Center and Department of Statistics, Pennsylvania State University. https://doi.org/10.1093/bib/bbz016

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

Gallagher, A. M., & Kaufman, J. C. (2005). *Gender differences in mathematics: What we know and what we need to know.* Cambridge University Press. https://psycnet.apa.org/record/2005-04568-015

Gilula, Z., & Haberman, S. J. (1994). Conditional log-linear models for analyzing categorical panel data. *Journal of the American Statistical Association, 89*, 645–656. https://doi.org/10.1080/01621459.1994.10476789

Holland, P. W., & Wainer, H. (1993). *Differential item functioning.* Lawrence Erlbaum. https://psycnet.apa.org/record/1993-97193-000

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, *104*(1), 53. https://doi.org/10.1037/0033-2909.104.1.53

Hyde, J. S., & Mertz, J. E. (2009). Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences*, *106*(22), 8801–8807. https://doi.org/10.1073/pnas.0901265106

Jewsbury, P. A., & van Rijn, P. W. (2019). IRT and MIRT Models for Item Parameter Estimation With Multidimensional Multistage Tests. *Journal of Educational and Behavioral Statistics*. https://doi.org/10.3102/1076998619881790

Kaufman, A. S., & Horn, J. L. (1996). Age changes on tests of fluid and crystallized ability for women and men on the Kaufman Adolescent and Adult Intelligence Test (KAIT) at

ages 17–94 years. *Archives of Clinical Neuropsychology*, *11*(2), 97–121. https://doi.org/10.1016/0887-6177(95)00003-8

Keith, T. Z., Reynolds, M. R., Patel, P. G., & Ridley, K. P. (2008). Sex differences in latent cognitive abilities ages 6 to 59: Evidence from the Woodcock–Johnson III tests of cognitive abilities. *Intelligence*, *36*(6), 502–525. https://doi.org/10.1016/j.intell.2007.11.001

Kennedy, A. M. (2008). Examining gender and fourth graders' reading habits and attitudes in PIRLS 2001 and 2006. *3rd IEA International Research Conference,* Taipei, Chinese Taipei.

Khorramdel, L., Shin, H. J., & von Davier, M. (2019). GDM software mdltm including parallel EM algorithm. In M. von Davier & Y-S. Lee (Eds.), *Handbook diagnostic classification models* (pp. 603–628). Springer. https://link.springer.com/chapter/10.1007%2F978-3-030-05584-4_30

Khorramdel, L., von Davier, M., Bertling, J. P., Roberts, R. D., & Kyllonen, P. C. (2017). Recent IRT approaches to test and correct for response styles in PISA background questionnaire data: A feasibility study. *Psychological Test and Assessment Modeling, 59,* 71-92. https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2017_20170323/05_Khorramdel.pdf

Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C., & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of a two solutions and three distractors item format to a one solution and five distractors item format. *International Journal of Selection and Assessment, 18*, 111–115. https://doi.org/10.1111/j.1468-2389.2010.00493.x

Lafontaine, D., & Monseur, C. (2009). Gender gap in comparative studies of reading comprehension: To what extent do the test characteristics make a difference? *European Educational Research Journal*, *8*(1), 69–79. https://doi.org/10.2304/eerj.2009.8.1.69

Lee, S. S., & von Davier, M. (2020). Improving measurement properties of the PISA home possessions scale through partial invariance modeling. *Psychological Test and Assessment Modeling, 62*, 55–83. https://www.psychologie-aktuell.com/fileadmin/Redaktion/Journale/ptam-2020-1/04_Lee.pdf

Lietz, P. (2006). A meta-analysis of gender differences in reading achievement at the secondary school level. *Studies in Educational Evaluation*, *32*(4), 317–344. https://doi.org/10.1016/j.stueduc.2006.10.002

Lindberg, S. M., Hyde, J. S., Petersen, J. L., & Linn, M. C. (2010). New trends in gender and mathematics performance: A meta-analysis. *Psychological Bulletin*, *136*(6), 1123–1135. https://doi.org/10.1037/a0021276

Logan, S., & Johnston, R. (2009). Gender differences in reading ability and attitudes: Examining where these differences lie. *Journal of Research in Reading*, *32*(2), 199–214. https://doi.org/10.1111/j.1467-9817.2008.01389.x

Maccoby, E. E., & Jacklin, C. N. (1978). *The psychology of sex differences* (Vol. 2). Stanford University Press. https://www.sup.org/books/title/?id=2885

Mackintosh, N. J. (1996). Sex differences and IQ. *Journal of Biosocial Science*, *28*(4), 558–571. https://doi.org/10.1017/S0021932000022586

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*(2), 149–174. https://doi.org/10.1007/BF02296272

Mau, W.-C., & Lynn, R. (2000). Gender differences in homework and test scores in mathematics, reading and science at tenth and twelfth grade. *Psychology, Evolution & Gender*, *2*(2), 119–125. https://doi.org/10.1080/14616660050200904

McGeown, S., Goodwin, H., Henderson, N., & Wright, P. (2012). Gender differences in reading motivation: Does sex or gender identity provide a better account? *Journal of Research in Reading*, *35*(3), 328–336. https://doi.org/10.1111/j.1467-9817.2010.01481.x

McKenna, M. C., Conradi, K., Lawrence, C., Jang, B. G., & Meyer, J. P. (2012). Reading attitudes of middle school students: Results of a US survey. *Reading Research Quarterly*, *47*(3), 283–306. https://www.jstor.org/stable/43497521

Mislevy, R. J., & Sheehan, K. M. (1987). Marginal estimation procedures. In A. E. Beaton (Ed.), *Implementing the new design: The NAEP 1983-84 technical report.* (Report No. 15-TR-20). Educational Testing Service. https://eric.ed.gov/?id=ED288887

Mücke, S. (2009). Schulleistungen von Jungen und Mädchen in der Grundschule – eine metaanalytische Bilanz. *Empirische Pädagogik*, *23*(3), 290–337. https://www.fachportal-paedagogik.de/literatur/vollanzeige.html?FId=888923

Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016: International results in reading.* https://timssandpirls.bc.edu/pirls2016/index-pirls.html

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159–177. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x

Nowell, A., & Hedges, L. V. (1998). Trends in gender differences in academic achievement from 1960 to 1994: An analysis of differences in mean, variance, and extreme scores. *Sex Roles*, *39*(1–2), 21–43. https://doi.org/10.1023/A:1018873615316

Organisation for Economic Co-operation and Development (2017a). Scaling PISA data. In OECD (Ed.), *PISA 2015 technical report* (pp. 128–185). OECD Publishing. https://www.oecd.org/pisa/data/2015-technical-report/09_Chapter_09_PISA2015.pdf

Organisation for Economic Co-operation and Development (2017b). Scaling outcomes. In OECD (Ed.), *PISA 2015 technical report* (pp. 128–185). OECD Publishing. http://www.oecd.org/pisa/data/PISA-2015-Technical-Report-Chapter-12-scaling.pdf

Organisation for Economic Co-operation and Development. (2019a). *PISA 2018 results (Volume II) | READ online*. OECD Publishing. https://read.oecd-ilibrary.org/education/pisa-2018-results-volume-ii_b5fd1b8f-en#page1

Organisation for Economic Co-operation and Development (2019b). PISA 2018 reading framework, in *PISA 2018 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/5c07e4f1-en

Organisation for Economic Co-operation and Development (2020a). Scaling PISA data. In OECD (Ed.), *PISA 2018 technical report*. OECD Publishing. https://www.oecd.org/pisa/data/pisa2018technicalreport/Ch.09-Scaling-PISA-Data.pdf

Organisation for Economic Co-operation and Development (2020b). Scaling outcomes. In OECD (Ed.), *PISA 2018 technical report*. OECD Publishing. https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018-TecReport-Ch-12%20-Scaling-Outcomes.pdf

Organisation for Economic Co-operation and Development & Statistics Canada. (2005). *Learning a living: First results of the Adult Literacy and Life Skills survey*. https://escholarship.org/uc/item/7nx5k5r9

Oliveri, M. E., & von Davier, M. (2011). Investigation of model fit and score scale comparability in international assessments. *Psychological Test and Assessment Modeling*, *53*, 315. https://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/04_Oliveri.pdf

Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing*, *14*, 1–21. ttps://doi.org/10.1080/15305058.2013.825265

O'Neil, H. F., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment*, *3*(2), 135–157. https://doi.org/10.1207/s15326977ea0302_2

Petscher, Y. (2010). A meta-analysis of the relationship between student attitudes towards reading and achievement in reading. *Journal of Research in Reading*, *33*(4), 335–355. https://doi.org/10.1111/j.1467-9817.2009.01418.x

Penk, C., & Richter, D. (2017). Change in test-taking motivation and its relationship to test performance in low-stakes assessments. *Educational Assessment, Evaluation and Accountability*, *29*(1), 55-79. https://doi.org/10.1007/s11092-016-9248-7

Rapp, J., & Borgonovi, F. (2019). Gender gap in mathematics and in reading: A within-student perspective. *Journal of Supranational Policies of Education (9)* 6–56. http://doi.org/10.15366/jospoe2019.9.001

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests.* Danish Institute for Educational Research. https://eric.ed.gov/?id=ED419814

Reckase, M. D. (2009) *Multidimensional item response theory (statistics for social and behavioral sciences.* Springer. https://www.springer.com/gp/book/9780387899756

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

Schwabe, F., McElvany, N., & Trendtel, M. (2015). The school age gender gap in reading achievement: Examining the influences of item format and intrinsic reading motivation. *Reading Research Quarterly*, *50*(2), 219–232. https://doi.org/10.1002/rrq.92

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 461–464. https://www.jstor.org/stable/2958889

Solheim, O. J., & Lundetræ, K. (2018). Can test construction account for varying gender differences in international reading achievement tests of children, adolescents and young adults? – A study based on Nordic results in PIRLS, PISA and PIAAC. *Assessment in Education: Principles, Policy & Practice*, *25*(1), 107–126. https://doi.org/10.1080/0969594X.2016.1239612

Stoet, G., & Geary, D. C. (2013). Sex differences in mathematics and reading achievement are inversely related: Within- and across-nation assessment of 10 years of PISA data. *PloS one*, *8*(3). https://doi.org/10.1371/journal.pone.0057988

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology*, *29*(1), 6–26. https://doi.org/10.1016/S0361-476X(02)00063-2

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation Matters: Using the Student Opinion Scale to Make Valid Inferences About Student Performance. *The Journal of General Education, 58(3), 129-151.* https://doi.org/10.1353/jge.0.0047

Tijmstra, J., Bolsinova, M., Liaw, Y. L., Rutkowski, L., & Rutkowski, D. (2019). Sensitivity of the RMSD for detecting item-level misfit in low-performing countries. *Journal of Educational Measurement.* https://doi.org/10.1111/jedm.12263

von Davier, M. (2005). Multidimensional discrete latent trait models (mdltm) [Computer software]. Educational Testing Service.

von Davier, M. (2010). Hierarchical mixtures of diagnostic models. *Psychological Test and Assessment Modeling, 52*, 8–28. https://www.psychologie-aktuell.com/fileadmin/download/ptam/1-2010/02_vonDavier.pdf

von Davier, M., Gonzalez, E., & Mislevy, R. (2009) What are plausible values and why are they useful? In *IERI Monograph Series: Issues and Methodologies in Large Scale Assessments, Vol. 2*. http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf

von Davier, M., Khorramdel, L., He, Q., Shin, H., & Chen, H. (2019). Developments in psychometric population models for technology-based large-scale assessments – An overview of challenges and opportunities. *Journal of Educational and Behavioral Statistics, 44*, 671–705. https://doi.org/10.3102/1076998619881789

von Davier, M., Rost, R., & Carstensen, C. H. (2007). Introduction: Extending the Rasch model. In M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 1-12). Springer.

von Davier, M. Sinharay, S., Oranje, A. & Beaton, A. (2006) Statistical procedures used in the National Assessment of Educational Progress (NAEP): Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics (Vol. 26): Psychometrics* (pp. 1039–1056). Elsevier.

von Davier, M., Yamamoto, K., Shin, H., Chen, H., Khorramdel, L., Weeks, J., Davis, S., Kong, N., & Kandathil, M. (2019). Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assessment in Education: Principles, Policy & Practice,* advance online publication. https://doi.org/10.1080/0969594X.2019.1586642

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450. https://doi.org/10.1007/BF02294627

Watson, K., Handal, B., & Maher, M. (2016). The influence of class size upon numeracy and literacy performance. *Quality Assurance in Education, (24)*4, 507–27. https://doi.org/10.1108/QAE-07-2014-0039

Willingham, W. W., & Cole, N. S. (2013). *Gender and fair assessment*. Routledge.

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., & 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

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*(3), 227–242. https://doi.org/10.1207/s15324818ame0803_3

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2018). Multistage Adaptive Testing Design in International Large-Scale Assessments. *Educational Measurement: Issues and Practice, 37,* 16-27. https://doi.org/10.1111/emip.12226

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2019). Introduction of Multistage Adaptive Testing Design in PISA 2018. *OECD Working Paper*. https://dx.doi.org/10.1787/b9435d4b-en

Supplemental Material – Tables

Table 7:
Effect Sizes (Cohen's d) of Mean Gender Differences Obtained from the 1D* Model for Each Country-by-Language-by-Gender Group

		Cohen's d	**Mean Gender Difference**	**Mean Score Reading**	
Country	*Language*	*Reading*	*Reading*	*Female*	*Male*
ALB	Albanian	0.47	38.31	421.15	382.84
ARE	English	0.40	44.92	498.61	453.69
ARE	Arabic	**0.82**	71.28	426.69	355.41
ARG	Spanish	0.20	19.65	413.92	394.27
AUS	English	0.27	28.89	519.01	490.12
AUT	German	0.27	26.93	497.39	470.46
BEL	German	0.29	27.08	499.49	472.42
BEL	Dutch	0.21	20.64	520.62	499.97
BEL	French	0.24	23.10	500.36	477.26
BGR	Bulgarian	0.40	42.29	437.91	395.61
BIH	Bosnian	0.33	28.32	408.20	379.89
BIH	Croatian	0.32	25.93	419.05	393.12
BIH	Serbian	0.44	35.52	416.19	380.67
BLR	Russian, Belarusian	0.24	22.06	487.73	465.67
BRA	Portuguese	0.24	24.32	421.76	397.45
BRN	English	0.29	27.31	426.94	399.64
CAN	English	0.26	25.45	538.11	512.67
CAN	French	0.32	29.51	528.69	499.18
CHE	German	0.29	30.66	491.51	460.85
CHE	Italian	0.37	31.77	511.35	479.58
CHE	French	0.31	28.76	515.70	486.94
CHL	Spanish	0.21	19.48	460.25	440.77
COL	Spanish	0.12	10.34	417.55	407.20
CRI	Spanish	0.17	13.89	432.48	418.59
CZE	Czech	0.32	31.20	506.69	475.49
DEU	German	0.23	23.81	511.38	487.57
DNK	Danish	0.32	29.09	518.52	489.43
DOM	Spanish	0.37	29.87	359.35	329.49
ESP	Spanish	0.25	22.82	489.75	466.93

		Cohen's d	Mean Gender Difference	Mean Score Reading	
Country	*Language*	*Reading*	*Reading*	*Female*	*Male*
ESP	Catalan	0.41	38.25	501.04	462.78
ESP	Basque	0.42	40.75	489.73	448.98
ESP	Galician	0.40	38.14	507.63	469.49
ESP	Valencian	0.32	28.59	462.04	433.45
EST	Estonian	0.38	33.28	552.19	518.91
EST	Russian	0.29	24.51	505.28	480.78
FIN	Finnish	**0.52**	49.17	546.35	497.18
FIN	Swedish	**0.69**	61.72	536.04	474.32
FRA	French	0.24	25.07	503.98	478.91
GBR	English	0.17	17.20	512.01	494.81
GBR	Welsh	0.36	30.84	448.71	417.87
GEO	Georgian, Russian	0.46	40.26	400.08	359.82
GEO	Azerbaijani	**0.53**	32.04	292.19	260.16
GRC	Greek	0.42	40.96	478.80	437.84
HKG	English	0.48	56.26	551.76	495.50
HKG	Chinese	0.36	33.88	546.59	512.70
HRV	Croatian	0.36	32.28	492.62	460.33
HUN	Hungarian	0.25	24.70	488.95	464.26
IDN	Indonesian	0.28	21.25	382.52	361.28
IRL	English, Irish	0.23	21.11	525.43	504.32
ISL	Icelandic	0.39	39.29	500.94	461.65
ISR	Hebrew	0.33	36.49	523.19	486.70
ISR	Arabic	**0.74**	70.47	398.55	328.08
ITA	German	0.02	2.45	496.13	493.68
ITA	Italian	0.24	23.69	490.32	466.63
JOR	Arabic	**0.65**	54.78	444.88	390.10
JPN	Japanese	0.21	20.96	513.21	492.25
KAZ	Kazakh	**0.52**	30.31	383.23	352.92
KAZ	Russian	0.21	17.98	451.41	433.43
KOR	Korean	0.22	22.38	533.86	511.48
KSV	Albanian	0.34	22.85	362.47	339.62
LBN	English	0.22	25.16	360.20	335.04
LBN	French	0.31	35.82	373.48	337.66
LTU	Polish	0.42	43.70	441.95	398.26

		Cohen's d	Mean Gender Difference	Mean Score Reading	
Country	*Language*	*Reading*	*Reading*	*Female*	*Male*
LTU	Lithuanian	0.42	39.95	496.45	456.51
LTU	Russian	0.30	26.81	487.49	460.68
LUX	German	0.35	36.52	488.19	451.68
LUX	English	0.16	13.06	586.27	573.21
LUX	French	0.17	18.77	463.36	444.58
LVA	Latvian	0.41	35.33	497.60	462.27
LVA	Russian	0.25	22.94	494.49	471.55
MAC	English	**0.52**	45.44	503.53	458.08
MAC	Chinese, Portuguese	0.18	15.02	549.44	534.43
MAR	Arabic	0.32	23.93	373.27	349.34
MDA	Romanian	0.47	41.96	434.92	392.97
MDA	Russian	0.45	37.69	488.72	451.03
MEX	Spanish	0.13	10.91	425.66	414.75
MKD	Albanian	0.46	39.79	370.27	330.48
MKD	Macedonian	**0.64**	56.21	437.85	381.64
MLT	English	0.43	48.64	471.94	423.30
MNE	Albanian	0.35	24.00	362.58	338.57
MNE	Serb (Yekavian)	0.37	32.54	437.12	404.59
MYS	Malay	0.32	24.87	424.17	399.30
MYS	English	0.19	19.17	461.67	442.50
NLD	Dutch	0.27	26.84	516.86	490.02
NOR	Bokmål	0.44	46.07	524.89	478.83
NOR	Nynorsk	0.41	41.00	502.19	461.19
NZL	English	0.26	28.08	521.43	493.36
PAN	Spanish, English	0.20	18.07	382.00	363.92
PER	Spanish	0.10	8.92	415.01	406.09
PHL	English	0.34	27.50	352.62	325.11
POL	Polish	0.35	33.38	529.16	495.79
PRT	Portuguese	0.22	21.49	502.22	480.73
QAT	English	0.35	37.68	483.21	445.54
QAT	Arabic	**1.00**	85.99	416.63	330.65
QAZ	Russian	0.20	18.73	450.28	431.55
QAZ	Azeri	0.38	27.44	391.10	363.66
QCI	Chinese	0.13	11.07	566.92	555.85

		Cohen's d	**Mean Gender Difference**	**Mean Score Reading**	
Country	*Language*	*Reading*	*Reading*	*Female*	*Male*
QCY	English	0.24	23.61	494.17	470.56
QCY	Greek	**0.53**	50.32	441.06	390.74
ROU	Romanian, Hungarian	0.36	34.78	445.59	410.81
RUS	Russian	0.28	26.45	495.39	468.94
SAU	Arabic, English	**0.72**	57.57	430.18	372.62
SGP	English	0.20	21.30	564.46	543.16
SRB	Serbian, Hungarian	0.37	36.87	455.98	419.10
SVK	Slovak	0.34	33.37	480.60	447.23
SVK	Hungarian	0.40	36.48	448.87	412.39
SVN	Slovenian	0.45	40.24	517.50	477.26
SWE	Swedish, English	0.31	33.17	524.77	491.60
TAP	Chinese	0.22	22.76	515.52	492.76
THA	Thai	0.49	38.77	417.95	379.18
TUR	Turkish	0.31	27.22	486.70	459.48
UKR	Russian	0.28	23.42	506.54	483.13
UKR	Ukrainian	0.39	35.24	482.95	447.71
URY	Spanish	0.24	23.09	439.67	416.58
USA	English	0.21	22.81	515.84	493.03
VNM	Vietnamese	0.31	22.23	515.06	492.84

Note: Medium to high Cohen's *d* values, that is, values larger than 0.5, are printed bold; the presented mean scores and mean gender differences are based on WLE_T values; a full table including *SD* and sample sizes (*N*) can be requested from the authors. Please see the full country names for the 3-letter country codes in Table 11.

Table 8:
Effect Sizes (Cohen's d) of Mean Gender Differences Obtained from the 2D-Source Model for each Country-by-Language-by-Gender Group

		Cohen's d		Mean Gender Difference		Mean Score Single		Mean Score Multiple	
Country	*Language*	*Single*	*Multiple*	*Single*	*Multiple*	*Female*	*Male*	*Female*	*Male*
ALB	Albanian	0.48	0.44	41.54	38.37	420.93	379.39	421.69	383.32
ARE	English	0.38	0.34	76.97	70.28	498.98	453.09	498.86	458.07
ARE	Arabic	**0.82**	**0.75**	45.89	40.79	426.04	349.07	426.23	355.95
ARG	Spanish	0.19	0.19	19.32	24.18	412.32	393.00	423.99	399.81
AUS	English	0.27	0.24	31.16	28.09	518.36	487.20	522.57	494.48
AUT	German	0.29	0.21	31.97	22.12	498.99	467.03	498.74	476.63
BEL	German	0.36	0.17	40.81	20.50	506.49	465.68	500.46	479.96
BEL	Dutch	0.21	0.17	26.98	21.16	520.50	498.15	521.38	504.03
BEL	French	0.27	0.21	22.35	17.35	497.22	470.24	504.15	482.99
BGR	Bulgarian	0.42	0.37	47.52	42.00	437.69	390.16	437.79	395.79
BIH	Bosnian	0.29	0.35	39.83	33.74	406.09	378.52	410.33	379.72
BIH	Croatian	0.31	0.26	27.75	22.24	415.97	388.22	420.03	397.79
BIH	Serbian	0.44	0.40	27.56	30.60	413.77	373.93	417.97	384.23
BLR	Russian, Belarusian	0.27	0.20	26.04	19.04	488.93	462.89	487.87	468.84
BRA	Portuguese	0.25	0.22	27.99	23.53	422.35	394.36	421.07	397.53
BRN	English	0.27	0.30	27.82	29.67	422.15	394.33	430.27	400.60
CAN	English	0.26	0.22	31.07	28.58	540.49	512.99	538.76	515.49
CAN	French	0.32	0.29	27.49	23.26	524.64	493.57	535.34	506.76
CHE	German	0.32	0.24	36.87	28.90	490.36	454.31	494.22	468.06
CHE	Italian	0.38	0.32	36.05	26.16	502.71	465.84	521.26	492.36
CHE	French	0.33	0.27	33.41	26.47	510.83	477.42	523.37	496.89
CHL	Spanish	0.22	0.18	21.74	18.23	461.34	439.60	460.44	442.21
COL	Spanish	0.13	0.10	12.19	9.93	417.53	405.34	418.46	408.52
CRI	Spanish	0.16	0.14	13.54	12.84	430.96	417.43	433.34	420.49
CZE	Czech	0.37	0.26	37.79	26.90	505.85	468.06	510.27	483.37
DEU	German	0.24	0.18	27.56	20.33	514.12	486.56	512.93	492.60
DNK	Danish	0.36	0.27	35.46	25.40	518.32	482.86	520.96	495.56
DOM	Spanish	0.39	0.32	33.08	27.48	357.35	324.27	358.44	330.96
ESP	Spanish	0.27	0.21	42.79	42.91	487.06	460.81	493.39	472.61
ESP	Catalan	0.39	0.36	39.52	35.87	498.43	458.92	503.20	467.33
ESP	Basque	0.41	0.42	37.16	36.05	491.10	448.31	491.58	448.67
ESP	Galician	0.38	0.34	29.81	26.88	502.02	464.85	513.54	477.50
ESP	Valencian	0.31	0.28	26.26	20.78	460.26	430.45	464.06	437.18
EST	Estonian	0.42	0.31	38.37	28.78	553.61	515.25	552.73	523.95

		Cohen's d		Mean Gender Difference		Mean Score Single		Mean Score Multiple	
Country	*Language*	*Single*	*Multiple*	*Single*	*Multiple*	*Female*	*Male*	*Female*	*Male*
EST	Russian	0.33	0.24	29.42	20.90	504.84	475.42	507.71	486.81
FIN	Finnish	**0.51**	0.45	63.20	59.63	550.29	496.94	545.85	501.62
FIN	Swedish	**0.70**	**0.62**	53.34	44.23	531.41	468.21	541.13	481.50
FRA	French	0.25	0.19	28.90	21.03	501.47	472.57	506.80	485.77
GBR	English	0.17	0.15	31.39	26.30	510.79	491.96	516.05	499.81
GBR	Welsh	0.28	0.26	18.83	16.25	447.05	415.66	453.22	426.92
GEO	Georgian, Russian	0.46	0.45	45.62	39.96	400.87	355.25	398.25	358.28
GEO	Azerbaijani	0.48	**0.54**	30.71	34.83	287.01	256.30	297.47	262.64
GRC	Greek	0.42	0.38	44.09	39.58	482.42	438.33	477.34	437.76
HKG	English	**0.52**	0.46	63.83	59.75	551.38	487.55	555.83	496.08
HKG	Chinese	0.32	0.33	32.25	33.84	548.07	515.82	547.10	513.27
HRV	Croatian	0.37	0.33	33.97	30.74	492.41	458.44	492.90	462.16
HUN	Hungarian	0.27	0.21	27.77	22.67	487.52	459.75	490.65	467.98
IDN	Indonesian	0.28	0.25	22.89	19.37	384.04	361.15	380.29	360.91
IRL	English, Irish	0.23	0.20	22.95	19.70	525.55	502.60	527.67	507.98
ISL	Icelandic	0.40	0.36	43.32	36.77	503.55	460.23	500.40	463.63
ISR	Hebrew	0.31	0.27	81.24	70.64	523.77	487.38	524.19	492.14
ISR	Arabic	**0.73**	**0.71**	36.39	32.05	400.16	318.91	396.02	325.39
ITA	German	-0.03	0.02	24.99	20.82	489.91	494.69	502.49	500.16
ITA	Italian	0.25	0.20	-4.78	2.34	489.43	464.44	492.25	471.43
JOR	Arabic	**0.64**	0.45	56.45	43.93	444.69	388.24	447.91	403.98
JPN	Japanese	0.18	0.19	19.70	20.29	511.25	491.55	516.09	495.80
KAZ	Kazakh	**0.53**	**0.50**	32.17	29.77	383.55	351.38	382.83	353.06
KAZ	Russian	0.26	0.15	23.88	13.07	452.36	428.47	451.71	438.65
KOR	Korean	0.23	0.18	26.04	18.97	535.46	509.42	536.08	517.11
KSV	Albanian	0.40	0.30	26.96	21.79	362.50	335.54	362.50	340.71
LBN	English	0.22	0.20	39.06	34.56	355.50	328.40	373.17	348.86
LBN	French	0.31	0.26	27.10	24.31	370.35	331.28	373.99	339.43
LTU	Polish	0.45	0.33	52.51	34.83	445.74	393.23	439.95	405.12
LTU	Lithuanian	0.42	0.37	42.30	37.43	497.99	455.69	496.00	458.57
LTU	Russian	0.37	0.21	35.91	21.66	490.71	454.80	488.71	467.06
LUX	German	0.35	0.29	40.07	32.07	486.60	446.53	489.99	457.92
LUX	English	0.12	0.15	21.90	19.18	583.16	573.69	589.85	576.79
LUX	French	0.18	0.16	9.47	13.06	456.64	434.73	469.04	449.86
LVA	Latvian	0.39	0.36	35.10	33.04	497.07	461.97	498.34	465.31
LVA	Russian	0.29	0.18	27.97	17.97	492.26	464.29	497.10	479.13
MAC	English	**0.53**	0.42	48.40	40.32	502.02	453.61	507.04	466.72

		Cohen's d		Mean Gender Difference		Mean Score Single		Mean Score Multiple	
Country	*Language*	*Single*	*Multiple*	*Single*	*Multiple*	*Female*	*Male*	*Female*	*Male*
MAC	Chinese, Portuguese	0.22	0.14	19.57	12.25	552.73	533.17	548.54	536.29
MAR	Arabic	0.34	0.29	26.65	23.01	374.00	347.36	371.45	348.44
MDA	Romanian	0.46	0.37	43.09	40.33	435.55	392.46	432.68	392.35
MDA	Russian	0.42	0.24	36.84	22.21	488.56	451.72	484.73	462.52
MEX	Spanish	0.15	0.09	13.45	8.14	426.64	413.20	424.75	416.61
MKD	Albanian	0.48	0.24	57.27	63.24	372.50	328.38	361.15	337.04
MKD	Macedonian	**0.63**	**0.56**	44.12	24.12	437.06	379.78	448.95	385.71
MLT	English	0.43	0.38	55.61	46.04	473.21	417.60	472.22	426.18
MNE	Albanian	0.39	0.31	36.18	29.91	363.00	335.10	365.65	343.91
MNE	Serb (Yekavian)	0.38	0.32	27.90	21.74	440.77	404.59	434.03	404.12
MYS	Malay	0.35	0.26	28.67	22.15	424.47	395.80	422.99	400.83
MYS	English	0.21	0.13	23.19	13.47	461.93	438.74	462.46	448.99
NLD	Dutch	0.28	0.23	29.09	24.21	517.74	488.65	518.51	494.31
NOR	Bokmål	0.46	0.41	54.49	32.38	528.13	476.64	526.85	482.24
NOR	Nynorsk	**0.52**	0.29	51.48	44.61	505.28	450.79	502.50	470.13
NZL	English	0.25	0.25	29.46	27.94	520.81	491.35	526.03	498.09
PAN	Spanish, English	0.21	0.19	19.37	19.18	381.02	361.65	381.70	362.53
PER	Spanish	0.11	0.10	10.75	9.42	414.00	403.24	416.92	407.50
PHL	English	0.32	0.34	27.22	28.56	347.44	320.22	353.34	324.78
POL	Polish	0.35	0.29	36.12	29.67	530.56	494.43	529.84	500.18
PRT	Portuguese	0.23	0.21	24.09	21.33	501.83	477.74	505.14	483.81
QAT	English	0.36	0.32	90.18	88.04	482.44	440.89	485.28	448.34
QAT	Arabic	**0.99**	**0.95**	41.55	36.94	415.36	325.18	417.97	329.93
QAZ	Russian	0.23	0.11	30.55	26.71	444.73	421.30	451.74	440.85
QAZ	Azeri	0.39	0.34	23.42	10.88	389.23	358.68	391.82	365.11
QCI	Chinese	0.13	0.11	12.06	9.62	565.72	553.66	570.29	560.67
QCY	English	0.26	0.23	57.09	48.74	496.48	467.16	495.61	472.74
QCY	Greek	**0.54**	0.48	29.33	22.87	442.10	385.01	439.81	391.07
ROU	Romanian, Hungarian	0.36	0.27	36.23	32.53	444.97	408.74	456.68	424.15
RUS	Russian	0.31	0.24	31.33	23.92	497.37	466.04	495.52	471.59
SAU	Arabic, English	**0.70**	**0.67**	59.44	54.14	429.76	370.32	434.88	380.73
SGP	English	0.17	0.19	19.27	21.33	566.62	547.35	566.71	545.39
SRB	Serbian, Hungarian	0.40	0.32	43.97	33.21	457.43	413.46	455.43	422.22
SVK	Slovak	0.34	0.30	39.24	34.17	477.39	441.18	484.62	453.90

		Cohen's d		Mean Gender Difference		Mean Score Single		Mean Score Multiple	
Country	*Language*	*Single*	*Multiple*	*Single*	*Multiple*	*Female*	*Male*	*Female*	*Male*
SVK	Hungarian	0.38	0.36	36.21	30.72	448.61	409.37	450.24	416.07
SVN	Slovenian	0.46	0.38	42.50	36.49	518.59	476.09	516.54	480.05
SWE	Swedish, English	0.35	0.26	39.38	30.27	524.92	485.54	529.06	498.79
TAP	Chinese	0.23	0.18	25.76	19.80	515.62	489.87	517.12	497.32
THA	Thai	0.47	0.48	39.55	39.63	415.24	375.69	420.52	380.89
TUR	Turkish	0.30	0.30	27.37	28.08	487.83	460.46	486.89	458.81
UKR	Russian	0.23	0.18	36.54	26.40	507.60	486.77	492.96	480.11
UKR	Ukrainian	0.38	0.28	20.83	12.85	483.04	446.50	484.08	457.69
URY	Spanish	0.26	0.20	28.07	20.88	438.62	410.55	440.52	419.65
USA	English	0.18	0.21	21.09	23.73	516.33	495.25	518.47	494.73
VNM	Vietnamese	0.32	0.15	23.91	12.23	515.88	491.97	515.89	503.67

Note: Medium to high Cohen's *d* values, that is, values larger than 0.5, are printed bold; the presented mean scores and mean gender differences are based on WLE_T values; a full table including *SD* and sample sizes (*N*) can be requested from the authors. Please see the full country names for the 3-letter country codes in Table 11.

Table 9:
Effect Sizes (Cohen's d) of Mean Gender Differences Obtained from the 3D-Format Model for Each Country-by-Language-by-Gender Group

		Cohen's d			Mean Gender Difference			Mean Score Continuous		Mean Score Non-Continuous		Mean Score Mixed	
Country	*Language*	*Continuous*	*Non-Continuous*	*Mixed*	*Continuous*	*Non-Continuous*	*Mixed*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*
ALB	Albanian	0.47	0.43	0.43	40.50	38.93	38.62	424.06	383.56	417.73	378.80	415.54	376.92
ARE	English	0.37	0.33	0.32	44.29	42.05	39.52	499.49	455.19	503.05	461.01	495.51	455.99
ARE	Arabic	**0.81**	**0.80**	**0.70**	76.74	72.93	66.35	430.92	354.18	419.27	346.33	417.15	350.80
ARG	Spanish	0.21	0.15	0.11	22.82	16.11	10.84	418.24	395.42	405.79	389.68	413.33	402.49
AUS	English	0.26	0.25	0.25	30.83	28.24	27.55	520.09	489.26	520.38	492.14	522.55	495.00
AUT	German	0.27	0.21	0.22	29.52	23.28	21.80	497.09	467.57	509.93	486.65	497.97	476.17
BEL	German	0.37	0.12	0.22	50.99	19.08	38.87	513.80	462.81	524.55	505.47	520.95	482.08
BEL	Dutch	0.20	0.17	0.16	21.85	16.89	17.07	522.62	500.77	522.24	505.35	518.21	501.15
BEL	French	0.24	0.18	0.24	24.70	17.45	24.84	500.11	475.42	498.69	481.24	504.91	480.06
BGR	Bulgarian	0.42	0.33	0.34	47.17	39.44	40.83	441.89	394.72	426.26	386.82	432.38	391.54
BIH	Bosnian	0.36	0.20	0.29	32.37	19.62	25.96	412.80	380.43	388.23	368.61	405.19	379.22
BIH	Croatian	0.34	0.29	0.15	28.57	27.76	14.00	425.02	396.45	397.19	369.43	408.88	394.88
BIH	Serbian	0.48	0.31	0.30	41.96	28.53	26.01	421.12	379.16	405.13	376.59	405.80	379.79
BLR	Russian, Belarusian	0.24	0.22	0.19	22.84	23.17	19.14	489.66	466.82	486.26	463.08	486.34	467.20
BRA	Portuguese	0.24	0.24	0.22	26.62	26.39	24.44	426.18	399.56	407.11	380.73	414.82	390.38
BRN	English	0.28	0.27	0.30	28.51	26.91	29.74	424.14	395.62	431.84	404.94	429.20	399.46
CAN	English	0.24	0.25	0.23	25.91	25.14	23.39	539.54	513.64	537.72	512.58	540.83	517.44
CAN	French	0.31	0.31	0.27	30.44	30.41	26.76	528.85	498.42	536.94	506.54	530.33	503.57
CHE	German	0.28	0.25	0.23	31.06	29.76	25.27	491.16	460.10	501.99	472.23	489.56	464.29
CHE	Italian	0.40	0.20	0.25	38.17	18.87	22.77	514.11	475.94	505.94	487.07	509.76	486.99

		Cohen's d			Mean Gender Difference			Mean Score Continuous		Mean Score Non-Continuous		Mean Score Mixed	
Country	*Language*	*Continuous*	*Non-Continuous*	*Mixed*	*Continuous*	*Non-Continuous*	*Mixed*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*
CHE	French	0.30	0.38	0.21	30.69	37.89	20.85	516.61	485.92	527.16	489.27	513.06	492.21
CHL	Spanish	0.19	0.19	0.22	19.80	17.33	22.03	463.05	443.25	451.96	434.63	460.35	438.31
COL	Spanish	0.11	0.13	0.12	10.42	13.44	11.03	418.72	408.30	418.64	405.21	414.72	403.69
CRI	Spanish	0.15	0.14	0.13	13.63	12.71	11.45	434.45	420.82	422.23	409.52	430.27	418.82
CZE	Czech	0.31	0.30	0.28	32.21	32.90	29.11	509.28	477.08	503.45	470.55	506.43	477.32
DEU	German	0.22	0.24	0.16	25.32	27.61	17.68	513.97	488.65	522.44	494.83	508.12	490.44
DNK	Danish	0.33	0.26	0.30	32.38	24.28	27.24	520.34	487.95	513.35	489.07	522.28	495.04
DOM	Spanish	0.36	0.24	0.39	29.98	20.82	34.09	360.70	330.71	342.07	321.25	356.24	322.15
ESP	Spanish	0.25	0.24	0.18	25.31	24.77	17.60	489.45	464.14	490.61	465.84	493.50	475.90
ESP	Catalan	0.37	0.34	0.38	39.64	30.66	35.94	499.95	460.30	506.80	476.14	502.95	467.02
ESP	Basque	0.39	0.41	0.29	40.78	62.44	46.64	484.88	444.10	523.12	460.67	495.14	448.49
ESP	Galician	0.36	0.22	0.35	38.31	36.67	32.69	507.99	469.68	518.83	482.16	507.64	474.96
ESP	Valencian	0.28	0.28	0.30	28.12	39.18	29.20	460.41	432.30	462.66	423.49	466.59	437.39
EST	Estonian	0.35	0.40	0.32	33.06	39.15	29.48	552.49	519.43	554.68	515.53	554.11	524.63
EST	Russian	0.31	0.22	0.24	28.23	19.22	22.47	505.61	477.39	507.45	488.23	509.23	486.75
FIN	Finnish	**0.51**	0.47	0.39	52.33	47.76	38.37	549.89	497.55	548.56	500.81	542.46	504.10
FIN	Swedish	**0.71**	**0.64**	**0.50**	68.71	51.59	50.38	537.23	468.52	533.25	481.66	542.96	492.58
FRA	French	0.24	0.19	0.17	26.92	21.55	19.65	503.50	476.58	511.28	489.73	503.44	483.79
GBR	English	0.17	0.14	0.15	18.63	14.28	15.33	512.05	493.42	518.74	504.46	515.50	500.17
GBR	Welsh	0.32	0.14	0.26	29.61	16.20	36.73	441.45	411.84	455.92	439.72	457.29	420.56
GEO	Georgian, Russian	0.49	0.36	0.39	45.76	37.48	36.15	403.84	358.08	389.56	352.08	390.91	354.75
GEO	Azerbaijani	**0.59**	0.49	0.24	36.46	63.75	17.34	299.31	262.85	260.89	197.14	286.62	269.28
GRC	Greek	0.41	0.36	0.36	43.07	36.74	38.08	482.95	439.89	475.40	438.65	472.23	434.15

		Cohen's d			Mean Gender Difference			Mean Score Continuous		Mean Score Non-Continuous		Mean Score Mixed	
Country	*Language*	*Continuous*	*Non-Continuous*	*Mixed*	*Continuous*	*Non-Continuous*	*Mixed*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*
HKG	English	**0.51**	0.42	0.34	63.62	61.06	43.18	550.19	486.57	554.53	493.47	566.81	523.63
HKG	Chinese	0.33	0.26	0.34	34.18	27.00	34.59	550.27	516.09	538.12	511.12	548.02	513.44
HRV	Croatian	0.38	0.30	0.27	35.27	29.77	25.47	495.81	460.54	489.57	459.80	486.45	460.98
HUN	Hungarian	0.27	0.15	0.21	27.57	14.87	23.48	492.17	464.61	478.30	463.43	487.93	464.46
IDN	Indonesian	0.28	0.18	0.25	22.10	15.39	20.47	383.06	360.96	379.53	364.13	378.14	357.66
IRL	English, Irish	0.22	0.25	0.17	22.39	22.78	16.25	525.92	503.53	524.82	502.04	530.16	513.92
ISL	Icelandic	0.39	0.32	0.35	41.97	35.38	36.56	502.97	461.00	497.52	462.15	500.88	464.32
ISR	Hebrew	0.28	0.31	0.34	34.06	38.15	37.87	525.94	491.88	521.92	483.77	523.57	485.70
ISR	Arabic	**0.73**	**0.74**	**0.63**	77.16	88.27	61.42	406.37	329.21	383.16	294.89	376.70	315.28
ITA	German	0.04	-0.06	-0.11	4.02	-12.22	-21.50	494.03	490.01	499.03	511.25	496.81	518.31
ITA	Italian	0.25	0.19	0.17	25.26	21.38	17.63	493.84	468.57	484.20	462.82	486.32	468.69
JOR	Arabic	**0.63**	**0.60**	0.48	56.31	55.55	42.17	444.63	388.32	445.39	389.84	444.87	402.69
JPN	Japanese	0.18	0.18	0.19	19.68	18.68	21.12	513.54	493.87	509.48	490.81	519.00	497.88
KAZ	Kazakh	**0.55**	0.43	0.45	31.69	27.85	30.92	384.02	352.32	373.05	345.20	385.74	354.82
KAZ	Russian	0.21	0.18	0.19	18.50	17.74	16.86	453.82	435.32	444.42	426.68	450.43	433.57
KOR	Korean	0.22	0.12	0.17	25.18	12.67	17.86	538.31	513.13	528.49	515.82	531.58	513.72
KSV	Albanian	0.40	0.18	0.27	26.98	13.23	21.47	366.62	339.64	351.13	337.90	355.41	333.94
LBN	English	0.25	0.17	0.08	31.18	22.18	8.63	358.23	327.05	350.92	328.74	372.62	363.99
LBN	French	0.31	0.27	0.35	40.54	35.22	38.36	370.78	330.25	365.23	330.01	393.81	355.45
LTU	Polish	0.38	0.30	0.27	43.44	43.91	28.84	445.99	402.55	435.85	391.93	429.57	400.73
LTU	Lithuanian	0.42	0.39	0.34	42.75	39.03	34.83	500.60	457.85	491.96	452.93	490.75	455.92
LTU	Russian	0.35	0.22	0.18	34.81	20.39	19.22	493.37	458.57	485.55	465.17	488.48	469.26
LUX	German	0.30	0.35	0.30	34.82	41.08	33.49	484.64	449.82	496.78	455.70	493.65	460.16
LUX	English	0.21	-0.13	0.08	18.39	-11.16	6.49	593.11	574.72	582.98	594.13	577.24	570.75

		Cohen's d			Mean Gender Difference			Mean Score Continuous		Mean Score Non-Continuous		Mean Score Mixed	
Country	*Language*	*Continuous*	*Non-Continuous*	*Mixed*	*Continuous*	*Non-Continuous*	*Mixed*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*
LUX	French	0.18	0.19	0.12	22.42	23.22	13.66	460.53	438.11	482.47	459.25	461.73	448.07
LVA	Latvian	0.38	0.40	0.36	34.82	33.96	33.49	499.71	464.89	492.70	458.74	495.64	462.14
LVA	Russian	0.25	0.18	0.22	24.32	17.96	22.15	493.72	469.40	495.96	478.00	497.80	475.64
MAC	English	0.49	0.47	0.36	47.89	40.32	31.82	502.75	454.86	519.08	478.76	501.27	469.45
MAC	Chinese, Portuguese	0.19	0.09	0.20	17.38	8.20	16.58	554.12	536.74	535.54	527.34	550.89	534.32
MAR	Arabic	0.34	0.25	0.26	26.98	20.89	20.21	375.97	348.99	369.17	348.28	363.74	343.53
MDA	Romanian	0.49	0.40	0.40	44.72	40.15	45.21	435.91	391.19	434.57	394.42	424.08	378.87
MDA	Russian	0.44	0.41	0.22	37.17	40.06	22.25	493.63	456.46	482.70	442.64	480.75	458.50
MEX	Spanish	0.13	0.10	0.12	11.35	8.82	10.41	426.82	415.47	425.67	416.85	422.14	411.74
MKD	Albanian	0.46	0.43	0.20	42.44	41.64	23.21	371.46	329.02	372.57	330.92	346.65	323.44
MKD	Macedonian	**0.65**	**0.57**	**0.59**	62.69	51.79	56.86	443.86	381.17	431.70	379.91	430.56	373.71
MLT	English	0.42	0.36	0.38	52.96	43.47	45.74	471.54	418.58	476.90	433.43	472.06	426.32
MNE	Albanian	0.47	-0.22	0.19	34.47	-14.94	14.32	372.99	338.51	327.17	342.11	354.88	340.57
MNE	Serb (Yekavian)	0.37	0.31	0.32	34.35	30.46	31.46	441.05	406.70	427.87	397.41	430.78	399.33
MYS	Malay	0.33	0.23	0.27	26.80	19.01	24.55	424.63	397.83	416.20	397.19	425.04	400.49
MYS	English	0.20	0.17	0.12	21.57	17.34	12.97	460.00	438.43	470.71	453.37	462.77	449.80
NLD	Dutch	0.26	0.24	0.23	27.91	24.57	24.41	518.27	490.37	517.71	493.13	518.75	494.33
NOR	Bokmål	0.42	0.47	0.40	48.23	49.80	44.27	527.00	478.77	531.13	481.33	528.73	484.46
NOR	Nynorsk	0.39	**0.59**	0.32	43.39	52.41	35.97	500.15	456.76	522.02	469.61	502.13	466.15
NZL	English	0.25	0.29	0.21	29.46	31.46	24.49	522.52	493.07	525.65	494.19	524.45	499.96
PAN	Spanish, English	0.22	0.16	0.18	20.64	15.79	18.29	383.49	362.85	373.27	357.49	378.82	360.52

		Cohen's d			Mean Gender Difference			Mean Score Continuous		Mean Score Non-Continuous		Mean Score Mixed	
Country	*Language*	*Continuous*	*Non-Continuous*	*Mixed*	*Continuous*	*Non-Continuous*	*Mixed*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*
PER	Spanish	0.12	0.10	0.05	11.65	9.45	4.80	419.68	408.04	408.59	399.14	405.60	400.80
PHL	English	0.34	0.29	0.33	27.72	25.03	30.64	351.40	323.67	355.16	330.13	346.38	315.74
POL	Polish	0.33	0.29	0.28	34.14	29.26	28.08	532.13	497.98	525.33	496.07	527.91	499.83
PRT	Portuguese	0.23	0.18	0.16	24.79	16.92	17.09	506.54	481.76	490.27	473.35	502.90	485.81
QAT	English	0.35	0.35	0.28	40.46	41.27	32.76	483.27	442.81	495.14	453.87	479.39	446.62
QAT	Arabic	**0.97**	**0.90**	**0.95**	89.50	88.35	89.24	420.68	331.18	406.02	317.67	410.25	321.01
QAZ	Russian	0.21	0.17	0.07	20.51	18.84	6.59	450.20	429.69	443.45	424.61	450.77	444.18
QAZ	Azeri	0.41	0.23	0.33	30.75	17.43	27.31	391.03	360.28	386.78	369.35	391.87	364.56
QCI	Chinese	0.12	0.03	0.16	11.07	3.39	14.25	569.54	558.46	561.29	557.90	570.15	555.90
QCY	English	0.24	0.32	0.20	26.09	30.25	20.27	495.24	469.15	506.47	476.22	492.82	472.55
QCY	Greek	**0.53**	0.48	0.44	55.22	53.60	47.05	443.32	388.09	440.33	386.73	434.52	387.48
ROU	Romanian, Hungarian	0.39	0.28	0.26	39.84	29.83	27.14	449.05	409.21	442.44	412.61	436.11	408.97
RUS	Russian	0.29	0.26	0.20	29.26	27.75	20.79	498.07	468.82	499.48	471.73	489.18	468.39
SAU	Arabic, English	**0.71**	**0.67**	**0.61**	58.44	58.73	53.37	429.71	371.27	431.32	372.59	425.33	371.96
SGP	English	0.16	0.14	0.21	19.62	15.91	23.62	566.14	546.52	566.00	550.09	568.97	545.36
SRB	Serbian, Hungarian	0.37	0.35	0.31	39.70	38.59	34.27	457.00	417.30	458.09	419.51	452.94	418.67
SVK	Slovak	0.34	0.23	0.32	35.34	24.35	34.40	482.98	447.64	468.83	444.48	484.91	450.51
SVK	Hungarian	0.31	0.42	0.43	30.31	37.07	48.00	448.47	418.16	444.87	407.80	456.54	408.55
SVN	Slovenian	0.41	0.42	0.40	39.67	39.86	39.07	518.45	478.78	517.59	477.73	515.53	476.46
SWE	Swedish, English	0.30	0.35	0.27	34.86	38.83	31.38	526.75	491.89	529.44	490.62	527.39	496.00

		Cohen's d			Mean Gender Difference			Mean Score Continuous		Mean Score Non-Continuous		Mean Score Mixed	
Country	*Language*	*Continuous*	*Non-Continuous*	*Mixed*	*Continuous*	*Non-Continuous*	*Mixed*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*
TAP	Chinese	0.20	0.19	0.23	22.46	21.45	24.70	516.29	493.83	511.98	490.53	520.02	495.32
THA	Thai	0.49	0.42	0.44	40.50	36.39	37.44	418.04	377.54	412.78	376.38	419.48	382.04
TUR	Turkish	0.30	0.26	0.29	28.18	22.25	28.34	490.43	462.25	474.77	452.52	485.43	457.09
UKR	Russian	0.31	0.13	0.20	26.24	12.28	20.10	516.15	489.91	484.18	471.90	514.45	494.35
UKR	Ukrainian	0.40	0.32	0.29	37.97	32.19	30.05	490.65	452.69	469.22	437.02	476.80	446.74
URY	Spanish	0.23	0.23	0.22	23.89	24.88	23.11	441.25	417.36	435.41	410.52	437.54	414.43
USA	English	0.19	0.21	0.16	23.23	24.53	18.05	518.13	494.90	517.54	493.01	515.86	497.81
VNM	Vietnamese	0.34	0.27	0.17	25.22	21.75	13.35	520.02	494.80	511.89	490.13	515.02	501.67

Note: Medium to high Cohen's *d* values, that is, values larger than 0.5, are printed bold; the presented mean scores and mean gender differences are based on WLE_T values; a full table including *SD* and sample sizes (*N*) can be requested from the authors. Please see the full country names for the 3-letter country codes in Table 11.

Table 10:
Effect Sizes (Cohen's d) of Mean Gender Differences Obtained from the 3D-Cognitive Model for each Country-by-Language-by-Gender Group

		Cohen's d			Mean Gender Difference			Mean Score Locate		Mean Score Understand		Mean Score Evaluate	
Country	*Language*	*Locate*	*Understand*	*Evaluate*	*Locate*	*Understand*	*Evaluate*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*
ALB	Albanian	0.40	0.48	0.43	33.50	41.95	38.89	410.48	376.97	425.74	383.79	420.87	381.98
ARE	English	0.37	0.33	0.41	43.11	40.18	48.38	495.06	451.96	494.34	454.16	511.25	462.87
ARE	Arabic	**0.75**	**0.79**	**0.78**	68.80	73.16	75.55	414.68	345.88	425.60	352.44	437.70	362.15
ARG	Spanish	0.15	0.18	0.23	16.01	18.82	24.62	410.21	394.20	411.47	392.65	419.54	394.92
AUS	English	0.25	0.26	0.24	27.55	29.92	28.72	516.33	488.78	517.96	488.04	528.00	499.28
AUT	German	0.26	0.27	0.19	27.85	28.98	20.22	498.49	470.64	500.15	471.18	495.31	475.08
BEL	German	0.44	0.25	0.10	66.15	38.45	14.44	524.53	458.38	517.59	479.13	477.60	463.16
BEL	Dutch	0.20	0.18	0.20	20.36	18.69	21.21	522.74	502.38	517.31	498.62	527.31	506.10
BEL	French	0.27	0.23	0.23	28.14	23.92	21.98	507.31	479.17	498.66	474.74	501.54	479.56
BGR	Bulgarian	0.38	0.41	0.34	43.20	46.19	38.91	436.01	392.81	439.23	393.04	436.35	397.44
BIH	Bosnian	0.32	0.31	0.37	30.95	28.12	30.97	407.74	376.79	411.58	383.45	401.23	370.27
BIH	Croatian	0.34	0.24	0.30	31.74	21.17	24.62	415.99	384.26	423.53	402.37	406.62	381.99
BIH	Serbian	0.36	0.42	0.48	32.10	38.00	38.21	415.52	383.42	420.83	382.84	406.14	367.93
BLR	Russian, Belarusian	0.17	0.24	0.24	16.67	23.27	23.53	489.19	472.52	490.24	466.97	483.98	460.45
BRA	Portuguese	0.24	0.24	0.23	27.32	25.92	25.20	412.71	385.39	421.84	395.92	428.83	403.64
BRN	English	0.29	0.28	0.25	31.11	27.89	25.79	433.33	402.22	424.74	396.86	423.34	397.55
CAN	English	0.25	0.26	0.21	24.10	27.19	22.26	534.52	510.42	538.62	511.43	544.29	522.04
CAN	French	0.34	0.29	0.29	31.27	29.10	28.87	529.26	497.99	524.86	495.76	540.60	511.73
CHE	German	0.29	0.27	0.22	32.18	31.04	24.78	495.09	462.91	493.30	462.26	486.32	461.54
CHE	Italian	0.32	0.29	0.45	29.57	26.41	43.44	497.00	467.43	509.52	483.11	527.82	484.38

		Cohen's d			Mean Gender Difference			Mean Score Locate		Mean Score Understand		Mean Score Evaluate	
Country	*Language*	*Locate*	*Understand*	*Evaluate*	*Locate*	*Understand*	*Evaluate*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*
CHE	French	0.29	0.34	0.24	28.16	33.86	23.60	516.74	488.59	518.52	484.66	515.74	492.14
CHL	Spanish	0.21	0.21	0.20	20.64	20.07	20.02	451.26	430.62	462.53	442.46	465.69	445.67
COL	Spanish	0.13	0.14	0.05	12.32	12.69	4.93	411.96	399.63	419.10	406.41	419.91	414.98
CRI	Spanish	0.14	0.15	0.13	12.40	13.33	12.59	430.59	418.19	433.58	420.25	426.57	413.97
CZE	Czech	0.30	0.34	0.25	31.61	35.00	24.83	511.38	479.77	509.06	474.05	504.64	479.82
DEU	German	0.28	0.20	0.17	31.76	23.05	19.14	519.79	488.04	512.59	489.54	511.69	492.55
DNK	Danish	0.37	0.31	0.27	34.95	30.18	24.85	522.34	487.39	517.12	486.93	522.87	498.02
DOM	Spanish	0.32	0.38	0.32	28.58	31.27	26.66	348.91	320.33	359.20	327.93	364.08	337.43
ESP	Spanish	0.26	0.23	0.24	25.60	22.26	23.86	487.79	462.18	488.97	466.71	496.01	472.15
ESP	Catalan	0.33	0.38	0.37	30.73	38.68	37.97	508.49	477.76	498.48	459.79	501.41	463.44
ESP	Basque	0.41	0.41	0.30	62.82	41.86	47.88	493.43	430.61	483.67	441.81	513.01	465.14
ESP	Galician	0.36	0.38	0.31	34.99	37.72	33.76	497.42	462.43	507.76	470.04	514.45	480.69
ESP	Valencian	0.32	0.28	0.32	33.71	25.03	35.11	478.90	445.18	453.68	428.65	466.10	430.99
EST	Estonian	0.35	0.38	0.30	30.79	36.17	29.23	557.88	527.09	554.18	518.01	547.30	518.07
EST	Russian	0.22	0.39	0.13	20.12	35.04	11.47	513.41	493.29	510.39	475.35	495.58	484.11
FIN	Finnish	**0.53**	0.47	0.46	51.69	49.03	45.80	555.79	504.10	547.59	498.56	542.63	496.83
FIN	Swedish	**0.55**	**0.64**	**0.69**	54.03	63.23	67.20	532.77	478.74	539.27	476.04	537.08	469.87
FRA	French	0.22	0.22	0.20	25.30	25.06	22.15	510.28	484.97	504.08	479.02	501.79	479.64
GBR	English	0.16	0.16	0.14	16.59	17.42	15.66	514.43	497.83	509.53	492.11	520.37	504.72
GBR	Welsh	0.25	0.11	0.24	29.65	13.03	29.91	458.51	428.86	438.04	425.01	459.95	430.04
GEO	Georgian, Russian	0.43	0.49	0.39	40.54	45.70	35.54	389.22	348.68	403.34	357.64	401.55	366.02
GEO	Azerbaijani	0.47	**0.62**	0.41	36.74	40.76	24.74	265.70	228.95	300.05	259.29	309.01	284.27
GRC	Greek	0.43	0.39	0.39	44.81	40.07	40.49	481.59	436.79	478.04	437.97	482.05	441.56
HKG	English	**0.55**	**0.51**	0.36	67.23	61.95	49.26	554.08	486.85	546.76	484.81	566.53	517.27

		Cohen's d			Mean Gender Difference			Mean Score Locate		Mean Score Understand		Mean Score Evaluate	
Country	*Language*	*Locate*	*Understand*	*Evaluate*	*Locate*	*Understand*	*Evaluate*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*
HKG	Chinese	0.26	0.30	0.39	26.27	30.59	39.15	543.96	517.69	546.51	515.92	552.09	512.94
HRV	Croatian	0.34	0.36	0.33	32.26	32.25	31.56	494.79	462.53	493.37	461.12	489.74	458.18
HUN	Hungarian	0.21	0.27	0.20	21.92	27.68	21.40	484.08	462.16	492.85	465.16	486.07	464.67
IDN	Indonesian	0.25	0.27	0.27	20.00	22.29	20.24	381.50	361.51	380.72	358.43	387.30	367.06
IRL	English, Irish	0.22	0.23	0.17	20.37	22.54	17.34	528.97	508.60	524.03	501.49	528.29	510.95
ISL	Icelandic	0.41	0.37	0.35	44.27	40.56	35.35	505.90	461.63	503.56	463.00	496.30	460.96
ISR	Hebrew	0.30	0.29	0.27	35.78	34.49	31.99	516.47	480.69	523.20	488.71	533.61	501.63
ISR	Arabic	**0.69**	**0.72**	**0.65**	72.92	74.36	68.87	385.43	312.51	399.55	325.19	403.33	334.46
ITA	German	-0.10	0.03	0.01	-20.10	4.10	1.25	484.16	504.26	503.36	499.27	488.33	487.08
ITA	Italian	0.24	0.20	0.24	25.48	20.80	24.61	484.19	458.70	491.10	470.29	495.74	471.13
JOR	Arabic	**0.54**	**0.61**	**0.70**	53.43	54.39	57.11	440.94	387.51	444.69	390.30	449.03	391.92
JPN	Japanese	0.20	0.18	0.18	20.43	19.81	20.32	510.91	490.48	515.41	495.60	512.23	491.91
KAZ	Kazakh	0.43	**0.52**	**0.56**	25.78	32.06	32.11	376.33	350.55	385.59	353.53	383.45	351.34
KAZ	Russian	0.18	0.24	0.14	16.13	21.31	12.26	455.08	438.95	457.31	436.00	438.32	426.06
KOR	Korean	0.20	0.21	0.19	21.07	22.65	21.70	532.35	511.28	537.40	514.75	533.27	511.57
KSV	Albanian	0.36	0.34	0.31	26.50	24.00	20.57	357.41	330.91	363.96	339.96	363.17	342.60
LBN	English	0.18	0.21	0.23	26.30	26.55	27.66	361.03	334.73	351.52	324.97	364.86	337.20
LBN	French	0.28	0.30	0.31	38.20	37.73	39.67	362.93	324.73	371.14	333.41	375.72	336.05
LTU	Polish	**0.51**	0.38	0.21	58.10	42.89	22.65	449.65	391.56	439.93	397.04	442.77	420.12
LTU	Lithuanian	0.39	0.43	0.33	37.89	43.96	32.90	494.03	456.14	499.97	456.01	493.66	460.76
LTU	Russian	0.23	0.36	0.24	20.94	35.14	24.66	489.07	468.13	492.67	457.53	486.47	461.81
LUX	German	0.34	0.33	0.26	37.23	37.36	29.85	489.15	451.92	490.15	452.78	482.61	452.76
LUX	English	0.35	0.08	0.19	23.00	6.54	18.13	583.49	560.49	583.90	577.36	601.18	583.05
LUX	French	0.17	0.19	0.11	19.77	22.51	13.82	471.08	451.30	463.17	440.66	456.25	442.43

		Cohen's d			Mean Gender Difference			Mean Score Locate		Mean Score Understand		Mean Score Evaluate	
Country	*Language*	*Locate*	*Understand*	*Evaluate*	*Locate*	*Understand*	*Evaluate*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*
LVA	Latvian	0.44	0.37	0.34	40.23	33.21	31.11	500.11	459.88	498.62	465.42	493.93	462.82
LVA	Russian	0.23	0.25	0.15	23.85	24.89	14.00	502.27	478.42	495.79	470.90	487.83	473.83
MAC	English	0.48	0.42	**0.53**	42.02	39.48	52.14	517.93	475.91	492.91	453.43	516.85	464.71
MAC	Chinese, Portuguese	0.11	0.16	0.24	9.90	14.38	22.72	544.12	534.22	551.38	537.00	554.02	531.30
MAR	Arabic	0.31	0.34	0.25	26.50	26.27	19.46	371.97	345.47	373.68	347.41	371.00	351.54
MDA	Romanian	0.43	0.43	0.49	41.77	41.27	46.17	438.76	397.00	436.13	394.86	429.69	383.52
MDA	Russian	0.43	0.44	0.29	44.31	38.77	24.38	491.44	447.13	492.34	453.57	480.71	456.33
MEX	Spanish	0.15	0.13	0.07	12.97	11.39	6.12	423.80	410.83	424.69	413.30	428.96	422.84
MKD	Albanian	0.41	0.47	0.42	43.41	43.20	38.42	378.99	335.59	372.71	329.51	361.05	322.63
MKD	Macedonian	**0.55**	**0.63**	**0.64**	51.76	59.54	60.43	443.60	391.84	438.05	378.50	433.35	372.93
MLT	English	0.41	0.40	0.38	51.23	50.37	46.36	481.07	429.84	469.62	419.24	472.57	426.21
MNE	Albanian	0.37	0.31	0.29	25.18	23.44	20.17	357.17	331.99	367.10	343.66	359.86	339.69
MNE	Serb (Yekavian)	0.34	0.35	0.38	32.82	33.36	32.93	437.66	404.83	438.67	405.31	436.18	403.25
MYS	Malay	0.31	0.34	0.20	28.22	27.46	16.48	428.06	399.84	422.88	395.41	421.86	405.38
MYS	English	0.26	0.11	0.23	28.12	12.10	24.23	484.08	455.96	453.56	441.46	464.63	440.40
NLD	Dutch	0.24	0.26	0.23	24.50	27.81	24.65	523.00	498.50	516.88	489.07	516.63	491.99
NOR	Bokmål	0.44	0.44	0.36	48.85	49.45	39.56	529.94	481.09	526.65	477.20	526.29	486.73
NOR	Nynorsk	0.44	0.42	0.34	43.81	45.61	40.18	509.39	465.58	500.87	455.26	503.33	463.16
NZL	English	0.26	0.23	0.26	29.11	25.97	31.17	523.92	494.81	521.06	495.09	526.43	495.26
PAN	Spanish, English	0.24	0.18	0.20	24.75	17.42	19.36	382.95	358.20	381.93	364.51	378.35	358.99
PER	Spanish	0.12	0.10	0.10	11.58	9.15	10.51	407.85	396.26	416.29	407.14	418.80	408.29
PHL	English	0.28	0.36	0.30	26.18	29.69	25.94	355.54	329.36	349.66	319.97	348.51	322.57

		Cohen's d			Mean Gender Difference			Mean Score Locate		Mean Score Understand		Mean Score Evaluate	
Country	*Language*	*Locate*	*Understand*	*Evaluate*	*Locate*	*Understand*	*Evaluate*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*
POL	Polish	0.37	0.32	0.25	36.92	32.84	25.69	531.60	494.68	530.35	497.51	527.50	501.81
PRT	Portuguese	0.16	0.23	0.24	16.03	23.87	24.66	500.53	484.50	501.94	478.06	509.46	484.80
QAT	English	0.30	0.34	0.32	34.10	39.02	38.20	478.69	444.59	481.39	442.38	492.28	454.08
QAT	Arabic	**0.94**	**0.96**	**0.98**	92.82	87.64	85.85	413.86	321.04	414.73	327.09	424.21	338.36
QAZ	Russian	0.17	0.21	0.15	18.52	21.15	13.85	453.46	434.94	453.20	432.05	439.50	425.65
QAZ	Azeri	0.32	0.37	0.42	24.10	28.94	31.97	387.33	363.23	393.13	364.19	388.87	356.89
QCI	Chinese	0.11	0.12	0.11	10.45	10.66	10.46	562.96	552.51	569.84	559.18	569.43	558.97
QCY	English	0.20	0.26	0.24	19.37	26.65	24.53	494.14	474.77	488.25	461.60	510.27	485.74
QCY	Greek	**0.53**	0.48	**0.52**	58.76	49.68	52.80	441.97	383.20	438.11	388.43	447.15	394.35
ROU	Romanian, Hungarian	0.36	0.33	0.35	38.00	33.56	37.01	450.71	412.71	439.53	405.96	454.11	417.10
RUS	Russian	0.28	0.28	0.23	28.88	28.50	22.11	498.46	469.59	498.65	470.15	491.08	468.97
SAU	Arabic, English	**0.66**	**0.69**	**0.73**	63.47	56.54	57.35	426.88	363.42	433.69	377.15	427.66	370.31
SGP	English	0.18	0.18	0.16	19.51	20.63	19.34	568.02	548.50	563.36	542.73	570.83	551.50
SRB	Serbian, Hungarian	0.39	0.36	0.32	42.54	38.96	32.38	456.60	414.06	458.95	419.98	451.53	419.15
SVK	Slovak	0.29	0.34	0.28	30.18	36.05	29.03	481.21	451.02	482.42	446.38	480.26	451.23
SVK	Hungarian	0.40	0.43	0.16	42.85	45.82	13.78	456.60	413.76	456.43	410.61	431.97	418.19
SVN	Slovenian	0.41	0.44	0.36	40.09	41.03	34.68	519.94	479.86	518.13	477.10	514.60	479.92
SWE	Swedish, English	0.35	0.29	0.29	39.62	33.34	33.74	532.36	492.74	523.61	490.27	529.95	496.21
TAP	Chinese	0.22	0.18	0.22	23.92	20.15	24.19	512.99	489.08	517.30	497.15	516.70	492.51
THA	Thai	0.46	**0.50**	0.41	38.50	41.56	33.31	412.98	374.48	420.69	379.13	415.44	382.13
TUR	Turkish	0.28	0.31	0.29	25.65	28.12	28.35	477.74	452.09	490.29	462.18	488.74	460.39

		Cohen's d			Mean Gender Difference			Mean Score Locate		Mean Score Understand		Mean Score Evaluate	
Country	*Language*	*Locate*	*Understand*	*Evaluate*	*Locate*	*Understand*	*Evaluate*	*Female*	*Male*	*Female*	*Male*	*Female*	*Male*
UKR	Russian	0.17	0.24	0.28	14.18	22.02	23.28	491.81	477.63	512.20	490.18	502.49	479.22
UKR	Ukrainian	0.37	0.37	0.35	35.94	35.97	33.41	477.22	441.27	486.44	450.47	481.01	447.60
URY	Spanish	0.26	0.24	0.17	26.98	24.98	17.86	434.24	407.26	443.07	418.09	437.65	419.79
USA	English	0.19	0.19	0.17	21.56	21.89	21.26	512.78	491.22	515.87	493.98	523.40	502.14
VNM	Vietnamese	0.30	0.30	0.29	23.54	22.44	20.90	509.01	485.47	518.88	496.44	516.58	495.67

Note: Medium to high Cohen's *d* values, that is, values larger than 0.5, are printed bold; the presented mean scores and mean gender differences are based on WLE_T values; a full table including *SD* and sample sizes (*N*) can be requested from the authors. Please see the full country names for the 3-letter country codes in Table 11.

Table 11:
Country Names for the 3-Letter Country Codes presented in Tables 7-10

Country Code	Country Name	Country Code	Country Name	Country Code	Country Name
ALB	Albania	HKG	Hong Kong (China)	NZL	New Zealand
ARE	United Arab Emirates	HRV	Croatia	PAN	Panama
ARG	Argentina	HUN	Hungary	PER	Peru
AUS	Australia	IDN	Indonesia	PHL	Philippines
AUT	Austria	ISL	Iceland	POL	Poland
BEL	Belgium	ISR	Israel	PRT	Portugal
BGR	Bulgaria	ITA	Italy	QAT	Qatar
BIH	Bosnia and Herzegovina	JOR	Jordan	QAZ	Baku (Azerbaijan)
BLR	Belarus	JPN	Japan	QCI	B-S-J-Z (China)
BRA	Brazil	KAZ	Kazakhstan	QCY	Cyprus
BRN	Brunei Darussalam	KOR	Korea	ROU	Romania
CAN	Canada	KSV	Kosovo	RUS	Russian Federation
CHE	Switzerland	LBN	Lebanon	SAU	Saudi Arabia
CHL	Chile	LTU	Lithuania	SGP	Singapore
COL	Colombia	LUX	Luxembourg	SRB	Serbia
CRI	Costa Rica	LVA	Latvia	SVK	Slovak Republic
CZE	Czech Republic	MAC	Macao	SVN	Slovenia
DNK	Denmark	MAR	Morocco	SWE	Sweden
DOM	Dominican Republic	MDA	Moldova	TAP	Chinese Taipei
ESP	Spain	MEX	Mexico	THA	Thailand
EST	Estonia	MKD	Macedonia	TUR	Turkey
FIN	Finland	MLT	Malta	UKR	Ukraine
FRA	France	MNE	Montenegro	URY	Uruguay
GBR	United Kingdom	MYS	Malaysia	USA	United States
GEO	Georgia	NLD	Netherlands	VNM	Vietnam
GRC	Greece	NOR	Norway		