Online Supplementary Materials

To Accompany

Gender Bias in Test Item Formats:

Evidence from PISA 2009, 2012, and 2015 Math and Reading Tests

Public State Data Analyses

Data analyzed for Colorado (CO) and Connecticut (CT) were obtained from publicly reported state aggregate data.[1]

Standardized mean differences for CO were estimated based on publicly reported scale score means and standard deviations for female and male students. The standardized mean difference was computed as

$$d = \frac{m_f - m_m}{\frac{sd_f + sd_m}{2}}.$$

Table S1 below reports the complete sample sizes, means, standard deviations and participation rates among CO students.

Results from CT were reported as the number of students scoring in each of four ordered proficiency categories for each test and did not include reported means and standard deviations. To estimate the standardized mean difference in test scores based on the ordered proficiency counts, heteroskedastic ordered probit (HETOP) models were estimated using the -hetop- package in Stata (Shear & Reardon, 2019) separately for each grade and subject. A HETOP model estimates the mean and standard deviation of a normal distribution underlying an observed set of proficiency counts for each group. The HETOP models assume that the cut scores used to determine proficiency are the same across gender groups within the same grade and test, but do not assume the cut scores are the same across different grade level tests. The means and standard deviations for male and female students were estimated under the constraint that the statewide distribution of scores within grade and subject were standardized with mean 0 and standard deviation of 1. A standardized mean difference in scores was estimated using the same equation above. Table S2 presents the complete proficiency counts for CT students.

[1] Data from Colorado were obtained at: https://www.cde.state.co.us/code/accountability-dataexplorertool. Data from Connecticut were obtained at: https://public-edsight.ct.gov/?language=en_US.

Table S1.

Colorado Aggregate Test Score Summary Statistics, by Grade, Year, and Subject

Year	Grade	Subject	Test	Gender	N	Part. Rate	Mean	SD
2018	8	ELA	CMAS	Female	28166	88.0%	754	38
2018	8	ELA	CMAS	Male	30066	89.6%	734	39
2018	8	Math	CMAS	Female	28323	87.9%	736	37
2018	8	Math	CMAS	Male	30276	89.6%	732	39
2019	9	EBRW	PSAT	Female	30708	93.9%	471	91
2019	9	EBRW	PSAT	Male	31839	93.0%	446	94
2019	9	Math	PSAT	Female	30774	93.9%	451	88
2019	9	Math	PSAT	Male	31905	93.0%	447	100

Table S2.

Connecticut Grade 11 Aggregate Test Score Proficiency Counts, by Year and Subject

						Frequency Counts			
Year	Test/Grade	Subject	Gender	Part. Rate	N	Level 1	Level 2	Level 3	Level 4
2015	SBAC	ELA	Female	81.2%	15814	2598	3710	5588	3918
2015	SBAC	ELA	Male	82.8%	16554	4569	4224	4853	2908
2015	SBAC	Math	Female	80.3%	15716	6805	3960	3251	1700
2015	SBAC	Math	Male	81.9%	16461	8118	3490	2818	2035
2016	SAT	ELA	Female	95.4%	18746	2849	3321	9249	3327
2016	SAT	ELA	Male	93.3%	19107	3836	3235	8836	3200
2016	SAT	Math	Female	95.4%	18732	4144	7497	5298	1793
2016	SAT	Math	Male	93.3%	19084	4611	6693	5366	2414

Item Content

Additional item features beyond those described in the text were reported in the PISA technical documentation for both subjects but were not reported consistently across all three years and hence were not included. The math process classifications were developed by content area experts for use with the 2012 and 2015 PISA administrations. Because 34 of the 35 items from 2009 were used in 2012, the classifications were applied to the 34 items used in 2009, while the 35th item in 2009 was removed. The regression analyses described in the main text combine the text format "mixed" and "multiple" categories due to the small number of items with "multiple" text formats. Table S3 presents the total number of items included in the analyses overall and by subscale and process (mathematics) or text format and process (reading).

Table S3.
Number of Items by Item Properties, Year, and Format

	2009		2012		2015	
Item Type	**CR**	**MC**	**CR**	**MC**	**CR**	**MC**
Math						
All Items	18	16	51	33	40	29
Change and Relationships	6	3	16	5	11	5
Quantity	5	5	11	10	8	10
Space and Shape	5	3	14	7	11	6
Uncertainty and data	2	5	10	11	10	8
Employ	10	4	22	14	16	13
Formulate	6	4	20	7	16	5
Interpret	2	8	9	12	8	11
Reading						
All Items	53	48	24	20	46	42
Access and retrieve	16	7	7	3	14	8
Integrate and interpret	17	36	9	15	15	31
Reflect and evaluate	20	5	8	2	17	3
Continuous	31	31	13	13	27	27
Mixed	4	7	2	2	2	5
Multiple	1	0	1	0	1	2
Non-continuous	17	10	8	5	16	8

Note. MC=multiple choice; CR=constructed response.

The PISA technical reports provide detailed item format categories. These categories varied across years. Table S4 shows how the detailed item format categories for each year were converted to the MC/CR dichotomy for analyses in the main paper.

Table S4.

Detailed Item Format Categories.

MC/CR	2009 Categories	2012 Categories	2015 Categories
MC	SMC CMC	SMC CMC	SMC - Computer Scored CMC - Computer Scored
CR	Closed CR Short response Open CR	CR Auto Coded CR Expert Coded CR Manual Coded	OR Computer Scored OR Human Coded

Note. MC=multiple choice; CR=constructed response; CMC=complex multiple choice; SMC=simple multiple choice; OR=open response.

Descriptive Statistics for International Samples

The following tables provide summary information about the 35 countries included in the international comparative analyses. Table S5 reports the 35 jurisdictions (countries) included and the test administration language used for the analyses. Table S6 summarizes the following characteristics of the data across the 35 countries in each year and subject: student sample sizes, number of items, average item p-values, standardized mean differences in percent correct scores, and item format effects (difference in standardized mean differences for CR versus MC item percent correct scores). Table S7 reports summary statistics for the combined sample pooling across all 35 countries using the same format as Table 2 of the main text. Finally, Figure S1 presents boxplots summarizing the distribution of linear logistic test model (LLTM) differential facet functioning (DFF) model parameter estimates across countries by subject and year.

Table S5.
List of Jurisdictions (Countries) and PISA Reported Test Languages

PISA Country Code (CNT)	Country/Jurisdiction	Test Language
AUS	Australia	English
AUT	Austria	German
BEL	Belgium	Dutch
CAN	Canada	English
CZE	Czech Republic	Czech
DEU	Germany	German
DNK	Denmark	Danish
ESP	Spain	Spanish
FIN	Finland	Finnish
FRA	France	French
GBR	United Kingdom	English
GRC	Greece	Greek
HKG	Hong Kong (China)	Chinese (Cantonese in 2012)
HRV	Croatia	Croatian
HUN	Hungary	Hungarian
IRL	Ireland	English
ISR	Israel	Hebrew
ITA	Italy	Italian
JPN	Japan	Japanese
KOR	Korea	Korean
LTU	Lithuania	Lithuanian
MNE	Montenegro	Serbian (Montenegrin in 2009)
NLD	Netherlands	Dutch
NOR	Norway	Norwegian (Bokmål in 2015)
POL	Poland	Polish
PRT	Portugal	Portuguese
RUS	Russia	Russian
SGP	Singapore	English
SVK	Slovak Republic	Slovak
SVN	Slovenia	Slovenian
SWE	Sweden	Swedish
TAP	Chinese Taipei (Taiwan)	Chinese (Mandarin in 2012)
THA	Thailand	Thai
TUR	Turkey	Turkish
USA	United States	English

Note. A small number of countries administered PISA in multiple languages within years. For the analyses in this paper only responses for students taking the test administered in the more

widely used language across the three years were included. For example, in Canada PISA is administered in both French and English, but more students participated in the English version.

Table S6.
Summary Statistics for Sample Sizes and Observed Score Statistics Across Countries.

	Math			Reading		
Variable	**2009**	**2012**	**2015**	**2009**	**2012**	**2015**
Students						
Mean	4851	7066	2820	7002	4876	2824
SD	3736	5416	1066	5403	3729	1066
Min	2677	3868	2048	3839	2641	2054
Max	20276	29357	6321	29314	20171	6340
Items						
Mean	33.89	83.91	68.80	100.29	43.86	87.86
SD	0.32	0.28	0.47	1.30	0.43	0.36
Min	33	83	67	95	42	87
Max	34	84	69	101	44	88
P Values						
Mean	0.48	0.48	0.46	0.59	0.59	0.59
SD	0.08	0.07	0.07	0.06	0.05	0.06
Min	0.27	0.29	0.31	0.41	0.43	0.44
Max	0.61	0.62	0.59	0.69	0.69	0.67
Cohen's d						
Mean	-0.11	-0.09	-0.10	0.34	0.31	0.27
SD	0.07	0.08	0.09	0.10	0.10	0.09
Min	-0.19	-0.22	-0.28	0.18	0.16	0.11
Max	0.05	0.13	0.11	0.58	0.52	0.50
Format Difference						
Mean	0.04	0.06	0.06	0.08	0.10	0.09
SD	0.05	0.04	0.04	0.04	0.05	0.04
Min	-0.06	0.00	-0.03	-0.03	0.00	0.01
Max	0.15	0.18	0.13	0.15	0.22	0.15

Note. Each variable is summarized across N=35 countries within subject and year. Standardized mean differences in percent correct scores are computed within each country as:

$$d_{all} = \frac{(m_f - m_m)}{\frac{sd_f + sd_m}{2}}$$

Where m_f and m_m and sd_f and sd_m are the means and standard deviations of percent correct scores based on all items a student was administered for female and male students, respectively. The standardized item format effects are calculated as:

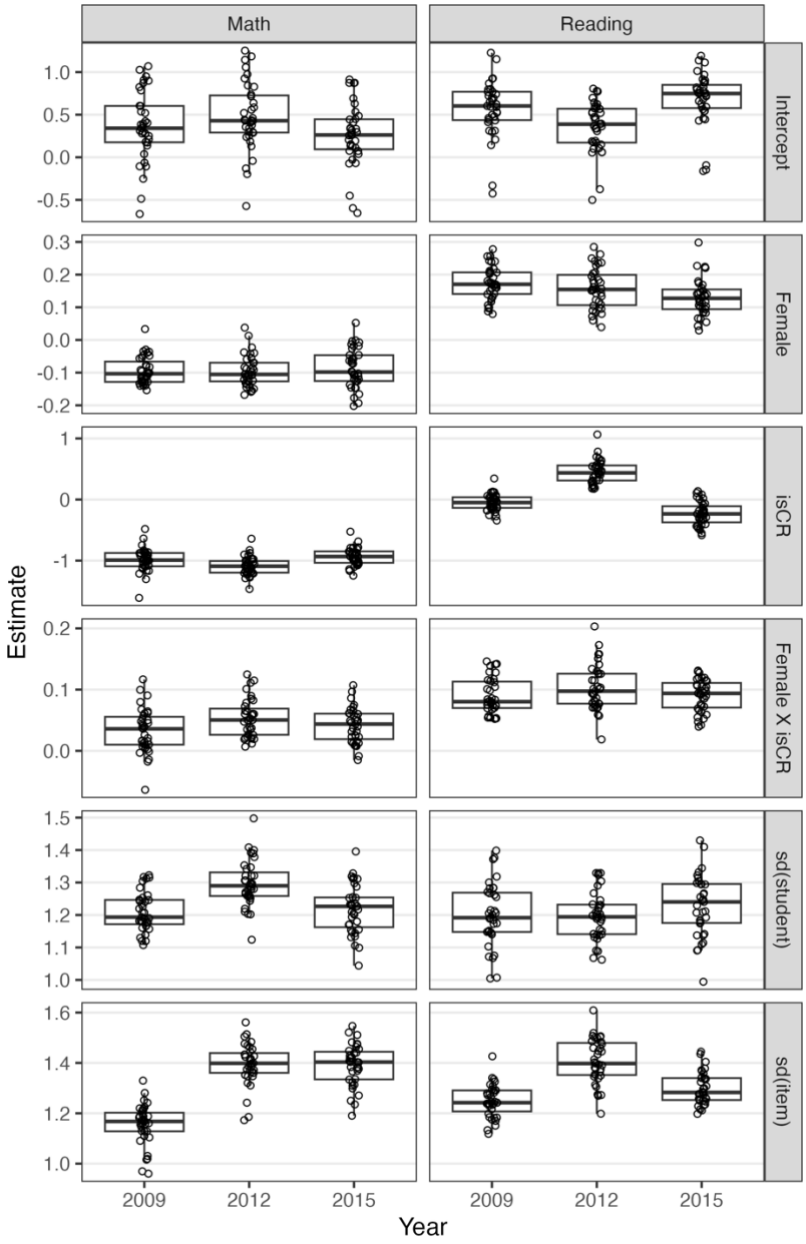
$$d_{format} = d_{CR} - d_{MC}$$

Where d_{CR} and d_{MC} are calculated analogously as d_{all}, but are based on percent correct scores for only CR items or MC items, respectively. The estimates of d_{CR} and d_{MC} are limited to students who responded to at least one CR and at least one MC item.

Table S7.
Combined Summary Table for Full International Sample

	Math			Reading		
	2009	**2012**	**2015**	**2009**	**2012**	**2015**
Number of Items						
All	34	84	69	101	44	88
CR	18	51	40	53	24	46
MC	16	33	29	48	20	42
Average P-Values						
All	0.47	0.48	0.46	0.59	0.59	0.59
CR	0.40	0.41	0.40	0.58	0.62	0.58
MC	0.55	0.58	0.54	0.59	0.55	0.61
Standardized Mean Differences						
All	-0.13	-0.11	-0.11	0.31	0.29	0.25
CR	-0.09	-0.08	-0.07	0.31	0.30	0.27
MC	-0.13	-0.13	-0.14	0.24	0.20	0.18
Format Diff.	0.03	0.05	0.06	0.07	0.10	0.09
Students						
N	169,795	247,323	98,690	245,055	170,670	98,838
% Female	50.0%	49.9%	49.8%	50.0%	49.7%	49.7%

Figure S1.
LLTM DFF Parameter Estimates Across Countries



Precision-Weighted Meta Analysis Regression Models

To account for the uncertainty of the differential item functioning (DIF) estimates analyzed in the linear regression models presented in the main text, the following table presents results of random effects meta-analytic regression models. These models use the same predictor variables and sample analyzed in the primary text but include an additional random error term reflecting the sampling error of the DIF estimates. The model estimated is

$$\hat{\delta}_{iy} = \alpha_0 + \alpha_1(isCR_i) + \boldsymbol{\omega X}_{iy} + \eta_y + e_{iy} + u_{iy}.$$

In this model $e_{iy} \sim N(0, \sigma^2)$ is a normally distributed error reflecting unexplained variation in true DIF estimates for each item, δ_{iy}, while $u_{iy} \sim N\left(0, \tau_{iy}^2\right)$ represents sampling error in the observed DIF estimates. In this model it is assumed that the DIF estimates are a combination of the true DIF for each item plus a random error term,

$$\hat{\delta}_{iy} = \delta_{iy} + u_{iy}.$$

The other variables are equivalent to those described in the main text. Table S8 presents the estimated parameters from these models. The models were estimated using restricted maximum likelihood implemented in the "metafor" package in R (Viechtbauer, 2010).

Table S8.
Meta-analytic Regression Model Results

	Math 1		**Math 2**		**Reading 1**		**Reading 2**	
Predictors	*Est.*	*SE*	*Est.*	*SE*	*Est.*	*SE*	*Est.*	*SE*
Overall	-0.06	0.05	-0.02	0.07	-0.06	0.03	-0.07	0.05
Is CR	0.09 *	0.04	0.13 **	0.04	0.11 **	0.04	0.09 *	0.04
factor(year)2012	0.00	0.05	0.00	0.05	-0.01	0.05	-0.01	0.05
factor(year)2015	0.01	0.06	0.01	0.05	-0.01	0.04	-0.01	0.04
Difficulty			-0.02	0.02			-0.07 ***	0.01
Quantity			0.01	0.05				
Space and Shape			-0.07	0.06				
Uncertainty and Data			0.00	0.06				
Formulate			-0.11 *	0.05				
Interpret			-0.04	0.05				
Integrate and Interpret							0.04	0.05
Reflect and Evaluate							0.08	0.05
Text Non-continuous							-0.04	0.04
Text Mixed/Multiple							-0.06	0.05
Observations	186		186		233		233	
R^2	0.021		0.082		0.038		0.161	

** p<0.05 ** p<0.01 *** p<0.001*

References

Shear, B. R., & Reardon, S. F. (2019). *HETOP: Stata module for estimating heteroskedastic ordered probit models with ordered frequency data* (Version 3). Statistical Software Components. https://ideas.repec.org/c/boc/bocode/s458287.html

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3). https://doi.org/10.18637/jss.v036.i03