

CHAPTER 13

Examining eTIMSS Country Differences Between eTIMSS Data and Bridge Data A Look at Country-Level Mode of Administration Effects

Matthias von Davier
Pierre Foy
Michael O. Martin
Ina V.S. Mullis

The TIMSS & PIRLS International Study Center made every effort to ensure a seamless transition from the TIMSS paper-and-pencil format to the new, computer based eTIMSS. A major priority in developing the TIMSS 2019 assessment was ensuring that the eTIMSS and paperTIMSS assessments measured the same mathematics and science constructs, using the same items as much as possible.

The TIMSS 2019 Bridge Between eTIMSS and paperTIMSS

The purpose of this chapter is to help each eTIMSS country understand how a comparison between its eTIMSS data and its bridge data can be used to study how the transition to eTIMSS may have affected its TIMSS 2019 achievement results.

Based on an item equivalence pilot in 2017 (Fishbein, Martin, Mullis, & Foy, 2018) that indicated a modest mode of administration effect, countries transitioning to eTIMSS included a bridge to paperTIMSS to control for this mode effect. To provide bridging data, eTIMSS countries administered the complete computer-based eTIMSS 2019 assessment as well as a smaller, paper-based version of the trend items. That is, eTIMSS countries re-administered their eight blocks of trend items from 2015 in paperTIMSS format. The bridge booklets were administered to an additional sample of 1,500 students, sampled from about one-third of the schools selected for the full eTIMSS sample. As a random sample from the same student population, the bridge sample taking the trend items in paperTIMSS format is randomly equivalent to the full eTIMSS sample. As such, the bridge data form an intermediate link between eTIMSS countries' computer-based data in 2019 and their paper-based data in 2015, as well as to the paperTIMSS countries in 2019.

Using the bridge data as a link between paperTIMSS and eTIMSS is an example of equivalent groups linking. Because students were randomly assigned to the new (eTIMSS) and old (paperTIMSS) formats of an assessment, the students taking the two formats could be expected to have the same underlying skills and knowledge. They differ only in that they were randomly assigned to different formats and are otherwise equivalent. Underlying this approach is the principle of randomization, one of the central principles of experimental design (Box, Hunter, & Hunter, 2005), which aims to ensure that observed differences in results of groups exposed to experimental treatments are due to the treatments themselves and not to pre-existing differences between the groups.

Item Equivalence Between eTIMSS and paperTIMSS

TIMSS 2019 paid particular attention in converting its paper trend items (items used in TIMSS 2015 and retained in the 2019 assessment) to the eTIMSS computer-based version to ensure as much comparability as possible between response modes. This resulted in a large number of trend items that were very similar in presentation and response format between the paper and the eTIMSS assessments. In developing new items, there was an effort to capitalize on the digital environment and produce more engaging item types.

Having a substantial percentage of equivalent items between paper and eTIMSS strengthens the validity and interpretability of achievement results based on linking the two modes and enhances the randomly equivalent groups design. The more similarity between the paper and computer-based items, the more achievement differences between them are likely to be due to a mode effect. Therefore, TIMSS 2019 devoted considerable effort to identifying items that were equivalent or invariant with respect to paper and eTIMSS format, in content and psychometric properties (see [Chapter 12](#); von Davier et al., 2019a,b).

Exhibit 13.1 shows the counts of equivalent and non-equivalent items in eTIMSS 2019 at fourth grade and eighth grade for mathematics and science. The percentage of equivalent trend items ranged from 80 to 91 percent across fourth and eighth grades for mathematics and science. Moreover, high percentages of all the eTIMSS items were equivalent—ranging from 72 to 87 percent. The equivalent items come from the following three categories defined by response types: multiple choice, keyboard, or number pad. As could be anticipated, somewhat higher percentages of the trend items were equivalent compared to the new items.

Exhibit 13.1: eTIMSS 2019 Achievement Items by Mode of Administration Equivalence**eTIMSS 2019 Fourth Grade Item Equivalence**

Item Type		Mathematics			Science		
		Trend	New	Total	Trend	New	Total
Equivalent Items	Multiple Choice Items	41	24	65	47	39	86
	Keyboard Items	3	3	6	39	22	61
	Number Pad Items	30	22	52	—	—	—
All Equivalent Items		74	49	123	86	61	147
All Non-Equivalent Items		18	30	48	9	13	22
All Items		92	79	171	95	74	169
Percentage of Equivalent Items		80%	62%	72%	91%	82%	87%

eTIMSS 2019 Eighth Grade Item Equivalence

Item Type		Mathematics			Science		
		Trend	New	Total	Trend	New	Total
Equivalent Items	Multiple Choice Items	60	26	86	58	44	102
	Keyboard Items	9	10	19	47	26	73
	Number Pad Items	33	29	62	2	1	3
All Equivalent Items		102	65	167	107	71	178
All Non-Equivalent Items		12	27	39	10	23	33
All Items		114	92	206	117	94	211
Percentage of Equivalent Items		89%	71%	81%	91%	76%	84%

Country-Level Differences in Average Percent Correct on TIMSS 2019 Trend Items by Mode of Administration

To help users of the TIMSS 2019 data gain an understanding of the effect of changing from paper TIMSS to eTIMSS, the analyses in this section compare average performance between the paper bridge and eTIMSS on the trend items. This approach provides a model for investigating country mode effects for different types of items or student groups and a useful avenue for beginning to explore a country's transition to eTIMSS. The computations are described in Appendix 13A and are relatively straightforward.

Although the approach could be applied to any group of items, the analyses below were restricted to the more than 80 percent of the trend items that were found to be invariant between the two modes (see Exhibit 13.1). These items are virtually identical except for administration mode, and the two samples were designed to be randomly equivalent. Apart from sampling differences and deviations from the sampling design that have caused some departure from this equivalence of comparison groups, the performance differences between the paper bridge and eTIMSS can be attributed to a mode effect.

Exhibits 13.2 through 13.5 show for each eTIMSS country average performance on the invariant trend items for the paper bridge and eTIMSS samples as well as the average across the countries. Exhibits 13.2 and 13.3 show the results for fourth grade mathematics and science. Exhibits 13.4 and 13.5 show the same for eighth grade mathematics and science.

Exhibit 13.2: eTIMSS 2019 Average Percent Correct on Paper Bridge and eTIMSS Invariant Items – Fourth Grade Mathematics

Country	Paper_INV	eTIMSS_INV
Austria	54.84 (0.94)	52.87 (0.56)
Canada	47.89 (1.05)	46.24 (0.49)
Chile	31.98 (0.97)	29.93 (0.52)
Chinese Taipei	71.94 (0.66)	69.89 (0.42)
Croatia	47.24 (0.90)	44.25 (0.53)
Czech Republic	50.96 (1.53)	52.01 (0.69)
Denmark	52.23 (0.96)	49.42 (0.58)
England	59.45 (1.31)	57.32 (0.82)
Finland	53.93 (0.87)	51.51 (0.55)
France	40.04 (0.90)	39.08 (0.74)
Georgia	48.13 (1.91)	39.84 (0.84)
Germany	50.97 (1.06)	47.75 (0.61)
Hong Kong SAR	72.90 (1.78)	69.45 (0.83)
Hungary	52.80 (1.31)	48.72 (0.69)
Italy	48.14 (1.31)	46.38 (0.72)
Korea, Rep. of	69.26 (0.65)	67.56 (0.57)
Lithuania	60.02 (0.70)	54.65 (0.80)
Netherlands	53.13 (1.09)	51.39 (0.62)
Norway (5)	55.52 (1.07)	54.00 (0.56)
Portugal	54.75 (1.18)	49.37 (0.75)
Qatar	36.32 (1.15)	33.93 (0.76)
Russian Federation	61.82 (1.01)	60.39 (0.85)
Singapore	76.91 (1.18)	74.98 (0.90)
Slovak Republic	47.68 (1.32)	45.41 (0.75)
Spain	45.18 (1.12)	42.36 (0.49)
Sweden	49.55 (1.65)	47.90 (0.80)
United Arab Emirates	46.10 (1.92)	41.09 (0.34)
United States	56.22 (1.18)	53.94 (0.70)
International Average	53.42 (0.23)	50.77 (0.13)

() Standard errors appear in parentheses.

Exhibit 13.3: eTIMSS 2019 Average Percent Correct on Paper Bridge and eTIMSS Invariant Items – Fourth Grade Science

Country	Paper_INV	eTIMSS_INV
Austria	48.60 (0.80)	48.11 (0.44)
Canada	48.71 (1.05)	48.50 (0.34)
Chile	39.07 (0.83)	38.47 (0.48)
Chinese Taipei	58.19 (0.62)	55.46 (0.34)
Croatia	50.83 (0.75)	49.46 (0.46)
Czech Republic	50.55 (1.60)	51.08 (0.45)
Denmark	48.85 (0.89)	47.86 (0.46)
England	55.33 (1.07)	53.51 (0.52)
Finland	56.64 (0.81)	56.32 (0.49)
France	40.93 (0.80)	41.04 (0.61)
Georgia	42.82 (1.45)	35.40 (0.62)
Germany	51.51 (1.05)	48.87 (0.56)
Hong Kong SAR	55.19 (1.55)	51.47 (0.74)
Hungary	53.49 (1.21)	51.37 (0.58)
Italy	47.87 (0.79)	46.16 (0.60)
Korea, Rep. of	66.45 (0.56)	65.10 (0.48)
Lithuania	54.52 (0.68)	52.85 (0.55)
Netherlands	48.30 (0.99)	47.66 (0.60)
Norway (5)	54.53 (0.70)	52.62 (0.49)
Portugal	47.03 (0.76)	43.52 (0.43)
Qatar	42.23 (1.50)	37.06 (0.71)
Russian Federation	61.53 (0.99)	59.61 (0.78)
Singapore	69.00 (1.12)	67.32 (0.76)
Slovak Republic	49.02 (0.95)	48.08 (0.64)
Spain	49.04 (0.97)	46.74 (0.40)
Sweden	51.52 (1.34)	53.06 (0.74)
United Arab Emirates	46.06 (1.63)	41.82 (0.34)
United States	54.33 (1.14)	52.76 (0.57)
International Average	51.51 (0.20)	49.69 (0.11)

() Standard errors appear in parentheses.

Exhibit 13.4: eTIMSS 2019 Average Percent Correct on Paper Bridge and eTIMSS Invariant Items – Eighth Grade Mathematics

Country	Paper_INV	eTIMSS_INV
Chile	26.94 (0.72)	26.58 (0.45)
Chinese Taipei	70.20 (1.15)	65.04 (0.56)
England	48.02 (1.63)	42.11 (1.16)
Georgia	32.78 (1.21)	30.43 (0.84)
Hong Kong SAR	62.45 (1.68)	57.16 (0.95)
Hungary	47.42 (1.84)	43.53 (0.71)
Israel	46.23 (1.65)	44.86 (1.02)
Italy	39.46 (0.73)	37.16 (0.59)
Korea, Rep. of	69.71 (0.75)	65.02 (0.55)
Lithuania	44.23 (1.26)	42.99 (0.67)
Malaysia	37.21 (2.04)	32.00 (0.55)
Norway (9)	43.85 (0.98)	40.83 (0.56)
Qatar	33.69 (1.19)	28.55 (0.80)
Russian Federation	52.34 (1.91)	48.21 (1.23)
Singapore	74.53 (1.57)	67.46 (1.06)
Sweden	44.71 (1.11)	40.48 (0.62)
Turkey	41.13 (1.51)	39.75 (0.84)
United Arab Emirates	39.88 (1.90)	34.59 (0.38)
United States	45.31 (1.70)	43.84 (1.01)
International Average	47.37 (0.33)	43.72 (0.18)

() Standard errors appear in parentheses.

Exhibit 13.5: eTIMSS 2019 Average Percent Correct on Paper Bridge and eTIMSS Invariant Items – Eighth Grade Science

Country	Paper_INV	eTIMSS_INV
Chile	36.21 (0.60)	35.03 (0.49)
Chinese Taipei	61.50 (0.96)	58.14 (0.44)
England	50.08 (1.03)	45.75 (0.88)
Georgia	33.50 (1.16)	32.75 (0.59)
Hong Kong SAR	50.41 (1.03)	43.10 (0.96)
Hungary	48.69 (1.35)	47.64 (0.56)
Israel	43.75 (1.27)	45.42 (0.83)
Italy	41.16 (0.82)	41.06 (0.53)
Korea, Rep. of	57.07 (0.77)	54.64 (0.49)
Lithuania	47.64 (0.88)	48.00 (0.64)
Malaysia	39.23 (1.86)	36.98 (0.52)
Norway (9)	43.57 (0.89)	41.15 (0.48)
Qatar	44.34 (1.02)	39.76 (0.82)
Russian Federation	52.27 (1.18)	50.38 (0.97)
Singapore	67.62 (1.26)	65.64 (0.78)
Sweden	49.21 (1.08)	47.56 (0.62)
Turkey	48.39 (1.23)	47.01 (0.76)
United Arab Emirates	44.84 (1.72)	40.80 (0.38)
United States	48.99 (1.16)	47.87 (0.82)
International Average	47.81 (0.27)	45.72 (0.16)

() Standard errors appear in parentheses.

Exhibit 13.6 shows the international average percent correct across countries for the invariant trend items for the paper bridge and eTIMSS as well as the difference between them, together with their standard errors. At both fourth and eighth grades, there was a small but significant average international difference favoring the paper bridge in each subject, with a smaller difference in science than mathematics. These international mode effects require an international adjustment for each subject and grade before country differences can be properly evaluated.

Exhibit 13.6: eTIMSS 2019 International Average Percent Correct on Paper Bridge and eTIMSS Invariant Items

Grade 4	Bridge	eTIMSS	Difference
Mathematics	53.42 (0.23)	50.77 (0.13)	2.65 (0.26) ▲
Science	51.51 (0.20)	49.69 (0.11)	1.82 (0.23) ▲
Grade 8	Bridge	eTIMSS	Difference
Mathematics	47.37 (0.33)	43.72 (0.18)	3.66 (0.38) ▲
Science	47.81 (0.27)	45.72 (0.16)	2.09 (0.31) ▲

() Standard errors appear in parentheses.

▲ indicates the bridge students performed significantly higher than the eTIMSS students ($\alpha = 0.05$).

Exhibits 13.7 and 13.8 show for fourth grade mathematics and science the country mode differences between the paper bridge and eTIMSS data, having adjusted for the average international differences. The country deviations from the international percent correct (difference between country average percent correct and the international average percent correct) are shown for the paper bridge and eTIMSS, together with their standard errors. For example, Austria's deviation for the bridge was 1.41 (0.94) and for eTIMSS was 2.10 (0.56). The relative difference for the country is the difference between the two deviations, e.g., -0.69 for Austria, which is not significant. The relative difference represents the country mode difference adjusted for the average international difference between modes (see Appendix 13A).

Exhibit 13.7: eTIMSS 2019 Country Deviations from International Average Percent Correct for Paper Bridge and eTIMSS Invariant Items and their Differences – Fourth Grade Mathematics

Country	Bridge	eTIMSS	Difference
Austria	1.41 (0.94)	2.10 (0.56)	-0.69 (1.09)
Canada	-5.54 (1.04)	-4.53 (0.49)	-1.01 (1.15)
Chile	-21.45 (0.96)	-20.85 (0.52)	-0.60 (1.09)
Chinese Taipei	18.51 (0.68)	19.12 (0.42)	-0.61 (0.80)
Croatia	-6.18 (0.90)	-6.52 (0.53)	0.34 (1.04)
Czech Republic	-2.46 (1.49)	1.23 (0.67)	-3.70 (1.63) ▽
Denmark	-1.20 (0.95)	-1.36 (0.58)	0.16 (1.12)
England	6.02 (1.29)	6.55 (0.80)	-0.52 (1.51)
Finland	0.50 (0.87)	0.74 (0.54)	-0.23 (1.02)
France	-13.38 (0.90)	-11.70 (0.72)	-1.69 (1.15)
Georgia	-5.29 (1.85)	-10.93 (0.82)	5.64 (2.03) ▲
Germany	-2.46 (1.05)	-3.02 (0.60)	0.56 (1.21)
Hong Kong SAR	19.47 (1.73)	18.68 (0.81)	0.80 (1.91)
Hungary	-0.62 (1.29)	-2.05 (0.68)	1.42 (1.45)
Italy	-5.28 (1.28)	-4.39 (0.71)	-0.89 (1.47)
Korea, Rep. of	15.83 (0.66)	16.79 (0.56)	-0.95 (0.87)
Lithuania	6.59 (0.71)	3.88 (0.78)	2.71 (1.06) ▲
Netherlands	-0.29 (1.08)	0.62 (0.61)	-0.91 (1.24)
Norway (5)	2.10 (1.06)	3.22 (0.55)	-1.13 (1.19)
Portugal	1.33 (1.16)	-1.41 (0.73)	2.73 (1.37) ▲
Qatar	-17.11 (1.13)	-16.84 (0.74)	-0.27 (1.35)
Russian Federation	8.40 (1.00)	9.62 (0.83)	-1.22 (1.30)
Singapore	23.48 (1.16)	24.21 (0.88)	-0.72 (1.46)
Slovak Republic	-5.75 (1.29)	-5.37 (0.74)	-0.38 (1.49)
Spain	-8.24 (1.10)	-8.41 (0.49)	0.17 (1.21)
Sweden	-3.87 (1.61)	-2.87 (0.78)	-1.00 (1.79)
United Arab Emirates	-7.32 (1.86)	-9.69 (0.36)	2.36 (1.90)
United States	2.79 (1.16)	3.17 (0.68)	-0.38 (1.35)

() Standard errors appear in parentheses.

▲ indicates the bridge students performed significantly higher than the eTIMSS students ($\alpha = 0.05$).

▽ indicates the bridge students performed significantly lower than the eTIMSS students ($\alpha = 0.05$).

Exhibit 13.8: TIMSS 2019 Country Deviations from International Average Percent Correct for Paper Bridge and eTIMSS Invariant Items and their Differences – Fourth Grade Science

Country	Bridge	eTIMSS	Difference
Austria	-2.91 (0.80)	-1.58 (0.44)	-1.33 (0.91)
Canada	-2.79 (1.03)	-1.19 (0.35)	-1.60 (1.09)
Chile	-12.43 (0.83)	-11.21 (0.47)	-1.22 (0.95)
Chinese Taipei	6.69 (0.63)	5.77 (0.35)	0.91 (0.72)
Croatia	-0.67 (0.75)	-0.23 (0.46)	-0.44 (0.88)
Czech Republic	-0.96 (1.55)	1.39 (0.45)	-2.35 (1.62)
Denmark	-2.66 (0.88)	-1.83 (0.46)	-0.83 (0.99)
England	3.83 (1.05)	3.82 (0.52)	0.01 (1.17)
Finland	5.14 (0.81)	6.63 (0.48)	-1.49 (0.94)
France	-10.57 (0.80)	-8.65 (0.60)	-1.92 (1.00)
Georgia	-8.69 (1.41)	-14.29 (0.61)	5.60 (1.53) ▲
Germany	0.01 (1.03)	-0.82 (0.55)	0.83 (1.17)
Hong Kong SAR	3.68 (1.51)	1.78 (0.72)	1.90 (1.67)
Hungary	1.98 (1.18)	1.69 (0.57)	0.30 (1.31)
Italy	-3.64 (0.79)	-3.53 (0.58)	-0.10 (0.98)
Korea, Rep. of	14.94 (0.57)	15.41 (0.47)	-0.47 (0.74)
Lithuania	3.02 (0.68)	3.16 (0.54)	-0.14 (0.87)
Netherlands	-3.21 (0.98)	-2.03 (0.59)	-1.18 (1.14)
Norway (5)	3.03 (0.70)	2.93 (0.48)	0.09 (0.85)
Portugal	-4.47 (0.76)	-6.17 (0.43)	1.69 (0.87)
Qatar	-9.28 (1.46)	-12.62 (0.70)	3.35 (1.61) ▲
Russian Federation	10.03 (0.97)	9.93 (0.76)	0.10 (1.23)
Singapore	17.49 (1.10)	17.63 (0.74)	-0.14 (1.32)
Slovak Republic	-2.49 (0.94)	-1.60 (0.63)	-0.89 (1.13)
Spain	-2.47 (0.96)	-2.95 (0.40)	0.48 (1.04)
Sweden	0.02 (1.30)	3.37 (0.72)	-3.35 (1.49) ▼
United Arab Emirates	-5.44 (1.58)	-7.87 (0.35)	2.43 (1.62)
United States	2.82 (1.12)	3.07 (0.56)	-0.24 (1.25)

() Standard errors appear in parentheses.

▲ indicates the bridge students performed significantly higher than the eTIMSS students ($\alpha = 0.05$).

▼ indicates the bridge students performed significantly lower than the eTIMSS students ($\alpha = 0.05$).

Exhibits 13.9 and 13.10 provide the relative differences in the percentage correct metric for eighth grade mathematics and science, respectively. Note that overall international differences were accounted for so that within country comparisons reflect the relative differences.

Exhibit 13.9: eTIMSS 2019 Country Deviations from International Average Percent Correct for Paper Bridge and eTIMSS Invariant Items and their Differences – Eighth Grade Mathematics

Country	Bridge	eTIMSS	Difference
Chile	-20.44 (0.76)	-17.13 (0.47)	-3.30 (0.89) ▽
Chinese Taipei	22.83 (1.09)	21.32 (0.53)	1.51 (1.27)
England	0.64 (1.54)	-1.61 (1.10)	2.25 (1.93)
Georgia	-14.59 (1.15)	-13.29 (0.80)	-1.30 (1.45)
Hong Kong SAR	15.07 (1.59)	13.44 (0.90)	1.63 (1.87)
Hungary	0.05 (1.74)	-0.18 (0.67)	0.23 (1.90)
Israel	-1.15 (1.56)	1.14 (0.97)	-2.29 (1.88)
Italy	-7.91 (0.69)	-6.55 (0.56)	-1.36 (0.97)
Korea, Rep. of	22.34 (0.70)	21.31 (0.52)	1.03 (0.96)
Lithuania	-3.15 (1.19)	-0.73 (0.63)	-2.42 (1.40)
Malaysia	-10.16 (1.93)	-11.71 (0.52)	1.55 (2.04)
Norway (9)	-3.52 (0.93)	-2.89 (0.53)	-0.64 (1.14)
Qatar	-13.68 (1.12)	-15.17 (0.75)	1.49 (1.41)
Russian Federation	4.97 (1.81)	4.50 (1.17)	0.47 (2.19)
Singapore	27.16 (1.49)	23.74 (1.01)	3.42 (1.83)
Sweden	-2.66 (1.05)	-3.23 (0.59)	0.57 (1.26)
Turkey	-6.24 (1.43)	-3.97 (0.80)	-2.28 (1.68)
United Arab Emirates	-7.50 (1.80)	-9.12 (0.36)	1.63 (1.87)
United States	-2.06 (1.61)	0.13 (0.95)	-2.19 (1.91)

() Standard errors appear in parentheses.

▲ indicates the bridge students performed significantly higher than the eTIMSS students ($\alpha = 0.05$).

▽ indicates the bridge students performed significantly lower than the eTIMSS students ($\alpha = 0.05$).

Exhibit 13.10: eTIMSS 2019 Country Deviations from International Average Percent Correct for Paper Bridge and eTIMSS Invariant Items and their Differences – Eighth Grade Science

Country	Bridge	eTIMSS	Difference
Chile	-11.60 (0.63)	-10.69 (0.49)	-0.91 (0.80)
Chinese Taipei	13.69 (0.91)	12.42 (0.41)	1.26 (1.05)
England	2.27 (0.97)	0.03 (0.84)	2.24 (1.32)
Georgia	-14.32 (1.09)	-12.97 (0.56)	-1.35 (1.27)
Hong Kong SAR	2.60 (0.98)	-2.62 (0.91)	5.22 (1.37) ▲
Hungary	0.88 (1.27)	1.92 (0.53)	-1.04 (1.41)
Israel	-4.06 (1.20)	-0.30 (0.79)	-3.76 (1.47) ▽
Italy	-6.66 (0.77)	-4.66 (0.51)	-1.99 (0.97) ▽
Korea, Rep. of	9.26 (0.73)	8.92 (0.46)	0.33 (0.91)
Lithuania	-0.17 (0.83)	2.28 (0.61)	-2.46 (1.07) ▽
Malaysia	-8.59 (1.76)	-8.74 (0.49)	0.15 (1.85)
Norway (9)	-4.24 (0.84)	-4.57 (0.45)	0.33 (1.00)
Qatar	-3.47 (0.96)	-5.96 (0.77)	2.49 (1.27)
Russian Federation	4.45 (1.12)	4.66 (0.92)	-0.21 (1.48)
Singapore	19.81 (1.19)	19.92 (0.74)	-0.12 (1.44)
Sweden	1.40 (1.02)	1.84 (0.59)	-0.44 (1.22)
Turkey	0.58 (1.16)	1.29 (0.72)	-0.72 (1.40)
United Arab Emirates	-2.98 (1.63)	-4.92 (0.36)	1.95 (1.69)
United States	1.17 (1.10)	2.15 (0.78)	-0.98 (1.38)

() Standard errors appear in parentheses.

▲ indicates the bridge students performed significantly higher than the eTIMSS students ($\alpha = 0.05$).

▽ indicates the bridge students performed significantly lower than the eTIMSS students ($\alpha = 0.05$).

Exhibits 13.7 through 13.10 provide a way to evaluate whether countries had positive or negative mode effects over and above the international effect. Although most differences were not statistically significant given their standard errors, there were some differences, mostly small. Also, when computing a large number of significance tests, some number (5% or so) will appear statistically significant just by means of random sampling variability rather than underlying mode differences.

The **estimated** differences observed when looking at a relatively large number of country-mean differences follow a statistical distribution around the ‘true’ differences. Some are smaller and some are larger, and should the exercise be repeated, and another set of bridge and eTIMSS samples collected,

a country's difference most likely would be slightly lower or slightly higher. This is the well-known 'regression to the mean' effect, whereby if one repeats a data collection the observations showing the most extreme estimates in the original study may not show as extreme estimates in the replication (e.g. Efron, 2011).

Country-Level Differences in TIMSS 2019 Average Scale Scores (Plausible Values) by Mode of Administration

As a consequence of the eTIMSS-paperTIMSS linking approach used in the achievement scaling, the eTIMSS scale scores and bridge scale scores are on the same TIMSS 2019 achievement scales and can be directly compared. However, the eTIMSS scale scores are based on all of the 2019 achievement items, while the bridge scale scores are based on only the trend items and were estimated from samples one-third the size of eTIMSS. For each grade and subject, the item parameters from the paperTIMSS trend scaling were applied to the eTIMSS data with a small constant adjustment to account for the average international difference (the international mode effect) between the paper and eTIMSS versions (see [Chapter 12](#); von Davier et al., 2019a,b). Thus, country differences in scale scores between the eTIMSS and paper bridge data are the result of country mode effects and sampling differences.

Exhibits 13.11 and 13.12 show average country differences between average eTIMSS and paper bridge scale scores for fourth grade mathematics and science.

Exhibit 13.11: eTIMSS 2019 Average Scale Scores for eTIMSS and Paper Bridge and their Differences – Fourth Grade Mathematics

Country	Bridge Average Score	eTIMSS Average Score	Difference
Austria	534 (3.3)	539 (2.0)	-5 (3.8)
Canada	512 (3.8)	512 (1.9)	0 (4.3)
Chile	436 (4.4)	441 (2.7)	-5 (5.1)
Chinese Taipei	603 (2.6)	599 (1.9)	4 (3.2)
Croatia	511 (3.4)	509 (2.2)	2 (4.0)
Czech Republic	519 (7.3)	533 (2.5)	-14 (7.7)
Denmark	528 (3.6)	525 (1.9)	3 (4.1)
England	553 (5.0)	556 (3.0)	-3 (5.9)
Finland	533 (3.4)	532 (2.3)	1 (4.1)
France	481 (3.7)	485 (3.0)	-4 (4.8)
Georgia	505 (8.0)	482 (3.7)	23 (8.8) ▲
Germany	519 (4.2)	521 (2.3)	-2 (4.8)
Hong Kong SAR	607 (7.9)	602 (3.3)	6 (8.5)
Hungary	530 (5.1)	523 (2.6)	7 (5.8)
Italy	511 (4.9)	515 (2.4)	-4 (5.5)
Korea, Rep. of	595 (2.5)	600 (2.2)	-5 (3.3)
Lithuania	547 (2.8)	542 (2.8)	5 (3.9)
Netherlands	528 (4.1)	538 (2.2)	-9 (4.6) ▼
Norway (5)	540 (3.9)	543 (2.2)	-2 (4.4)
Portugal	536 (4.5)	525 (2.6)	11 (5.2) ▲
Qatar	450 (6.4)	449 (3.4)	0 (7.2)
Russian Federation	559 (3.9)	567 (3.3)	-8 (5.1)
Singapore	631 (5.6)	625 (3.9)	6 (6.8)
Slovak Republic	505 (4.7)	510 (3.5)	-5 (5.9)
Spain	502 (4.8)	502 (2.1)	-1 (5.2)
Sweden	517 (5.8)	521 (2.8)	-5 (6.4)
United Arab Emirates	496 (7.9)	481 (1.7)	14 (8.1)
United States	537 (5.1)	535 (2.5)	2 (5.7)
International Average	529 (1.0)	528 (0.6)	1 (1.2)

() Standard errors appear in parentheses.

▲ indicates the bridge students performed significantly higher than the eTIMSS students ($\alpha=0.05$).

▼ indicates the bridge students performed significantly lower than the eTIMSS students ($\alpha=0.05$).

Exhibit 13.12: eTIMSS 2019 Average Scale Scores for eTIMSS and Paper Bridge and their Differences – Fourth Grade Science

Country	Bridge Average Score	eTIMSS Average Score	Difference
Austria	511 (3.9)	522 (2.6)	-11 (4.7) ▽
Canada	512 (4.5)	523 (1.9)	-11 (4.9) ▽
Chile	461 (4.5)	469 (2.6)	-8 (5.1)
Chinese Taipei	554 (2.9)	558 (1.8)	-4 (3.4)
Croatia	524 (3.5)	524 (2.2)	0 (4.1)
Czech Republic	517 (9.4)	534 (2.6)	-16 (9.8)
Denmark	514 (4.3)	522 (2.4)	-8 (4.9)
England	543 (4.7)	537 (2.7)	6 (5.4)
Finland	547 (4.0)	555 (2.6)	-8 (4.7)
France	478 (4.0)	488 (3.0)	-10 (4.9)
Georgia	477 (8.1)	454 (3.9)	23 (9.0) ▲
Germany	522 (4.7)	518 (2.2)	4 (5.2)
Hong Kong SAR	542 (7.3)	531 (3.3)	11 (8.0)
Hungary	533 (6.3)	529 (2.7)	3 (6.8)
Italy	507 (4.1)	510 (3.0)	-3 (5.1)
Korea, Rep. of	588 (2.6)	588 (2.1)	0 (3.4)
Lithuania	539 (3.1)	538 (2.5)	1 (4.0)
Netherlands	511 (4.5)	518 (2.9)	-7 (5.3)
Norway (5)	536 (3.5)	539 (2.2)	-3 (4.1)
Portugal	509 (3.5)	504 (2.6)	5 (4.3)
Qatar	463 (8.6)	449 (3.9)	14 (9.5)
Russian Federation	567 (4.2)	567 (3.0)	0 (5.1)
Singapore	599 (5.1)	595 (3.4)	5 (6.2)
Slovak Republic	512 (4.9)	521 (3.7)	-9 (6.1)
Spain	514 (4.3)	511 (2.0)	3 (4.8)
Sweden	523 (6.4)	537 (3.3)	-15 (7.2) ▽
United Arab Emirates	485 (8.6)	473 (2.1)	12 (8.8)
United States	535 (5.4)	539 (2.7)	-3 (6.1)
International Average	522 (1.2)	523 (1.0)	0 (1.6)

() Standard errors appear in parentheses.

▲ indicates the bridge students performed significantly higher than the eTIMSS students ($\alpha = 0.05$).

▽ indicates the bridge students performed significantly lower than the eTIMSS students ($\alpha = 0.05$).

Exhibits 13.11 and 13.12 show that the differences for grade 4 are mostly non-significant. The only country with significant results in both mathematics and science was Georgia, where students taking eTIMSS had lower achievement than those taking the paper bridge. In Portugal, the students taking the eTIMSS appear to perform lower in mathematics compared to the bridge but not in science. In the Netherlands, students taking eTIMSS performed better in mathematics, but there was no difference in science. In Sweden, Austria, and Canada, the students taking eTIMSS performed better in science than those who took the bridge, while there was no difference in mathematics.

Exhibits 13.13 and 13.14 show for eighth grade mathematics and science the comparisons of bridge and eTIMSS samples in terms of achievement estimates (plausible values) based on the randomly equivalent groups.

Exhibit 13.13: eTIMSS 2019 Average Scale Scores for eTIMSS and Paper Bridge and their Differences – Eighth Grade Mathematics

Country	Bridge Average Score	eTIMSS Average Score	Difference
Chile	434 (3.3)	441 (2.8)	-6 (4.3)
Chinese Taipei	618 (5.4)	612 (2.7)	5 (6.1)
England	526 (6.0)	515 (5.3)	11 (8.0)
Georgia	452 (7.1)	461 (4.3)	-9 (8.3)
Hong Kong SAR	581 (6.9)	578 (4.1)	3 (8.0)
Hungary	521 (7.1)	517 (2.9)	5 (7.7)
Israel	511 (7.1)	519 (4.3)	-8 (8.3)
Italy	495 (3.5)	497 (2.7)	-2 (4.4)
Korea, Rep. of	613 (3.6)	607 (2.8)	7 (4.6)
Lithuania	510 (5.2)	520 (2.9)	-11 (6.0)
Malaysia	473 (9.9)	461 (3.2)	13 (10.4)
Norway (9)	509 (3.9)	503 (2.4)	7 (4.6)
Qatar	452 (6.3)	443 (4.0)	9 (7.5)
Russian Federation	543 (7.5)	543 (4.5)	-1 (8.8)
Singapore	630 (6.5)	616 (4.0)	15 (7.7)
Sweden	513 (4.8)	503 (2.5)	11 (5.5) ▲
Turkey	487 (7.1)	496 (4.3)	-9 (8.3)
United Arab Emirates	482 (8.6)	473 (1.9)	8 (8.8)
United States	512 (6.4)	515 (4.8)	-4 (8.0)
International Average	519 (1.5)	517 (0.8)	2 (1.7)

() Standard errors appear in parentheses.

▲ indicates the bridge students performed significantly higher than the eTIMSS students ($\alpha = 0.05$).

▽ indicates the bridge students performed significantly lower than the eTIMSS students ($\alpha = 0.05$).

Exhibit 13.14: eTIMSS 2019 Average Scale Scores for eTIMSS and Paper Bridge and their Differences – Eighth Grade Science

Country	Bridge Average Score	eTIMSS Average Score	Difference
Chile	458 (3.7)	462 (2.9)	-4 (4.7)
Chinese Taipei	584 (5.0)	574 (1.9)	10 (5.3)
England	529 (5.0)	517 (4.8)	13 (7.0)
Georgia	435 (7.5)	447 (3.9)	-12 (8.4)
Hong Kong SAR	531 (5.3)	504 (5.2)	27 (7.4) ▲
Hungary	523 (6.5)	530 (2.6)	-6 (7.0)
Israel	498 (7.0)	513 (4.2)	-16 (8.2)
Italy	487 (4.3)	500 (2.6)	-13 (5.0) ▼
Korea, Rep. of	563 (3.6)	561 (2.1)	2 (4.2)
Lithuania	522 (4.5)	534 (3.0)	-12 (5.4) ▼
Malaysia	469 (10.9)	460 (3.5)	9 (11.4)
Norway (9)	500 (4.7)	495 (3.1)	5 (5.6)
Qatar	495 (5.1)	475 (4.4)	20 (6.7) ▲
Russian Federation	544 (6.1)	543 (4.2)	1 (7.4)
Singapore	611 (6.1)	608 (3.9)	3 (7.3)
Sweden	521 (6.1)	521 (3.2)	0 (6.8)
Turkey	518 (6.4)	515 (3.7)	2 (7.4)
United Arab Emirates	490 (9.9)	473 (2.2)	17 (10.2)
United States	524 (6.0)	522 (4.7)	1 (7.6)
International Average	516 (1.6)	513 (0.8)	3 (1.8)

() Standard errors appear in parentheses.

▲ indicates the bridge students performed significantly higher than the eTIMSS students ($\alpha = 0.05$).

▼ indicates the bridge students performed significantly lower than the eTIMSS students ($\alpha = 0.05$).

As shown in Exhibits 13.13 and 13.14, the differences for the eighth grade are mostly non-significant. In Sweden, students taking eTIMSS performed lower in mathematics, but there was no difference in science. In Hong Kong SAR and Qatar, students taking eTIMSS performed lower in science than those who took the bridge, while there was no difference in mathematics. In Italy and Lithuania, the eTIMSS students performed higher in science than the bridge students.

Exhibits 13.11 through 13.14 show a small number of significant differences between eTIMSS and the paper bridge. Of the few countries with mode differences, they were either in mathematics or in science and occurred in both directions, with the exception of a paper bridge advantage for Georgia at the

fourth grade. Also, these exhibits contain a total of 94 mode comparisons, 56 for fourth grade and 38 for eighth grade, and we have not made adjustments for multiple comparisons (e.g. Shaffer, 1995; Benjamini & Hochberg, 1995). About 5 significant differences would be expected to occur purely at random among 94 comparisons at the $\alpha=0.05$ level.

Summary

The present chapter provides an overview of how countries can use their bridge data together with their eTIMSS data to evaluate the extent of mode differences in their TIMSS 2019 data. First, to provide an accessible approach to the study of country mode effects, country differences in the average percent correct between the paper bridge and eTIMSS were examined for those trend items found to be equivalent. Then, after subtracting the average international difference from the country average for both the bridge and eTIMSS, the difference between the bridge and eTIMSS country deviations provides an estimate of the country mode effect. However, only a few countries had significant mode effects, and these were for most countries isolated instances of one subject in one grade.

Second, country differences between average eTIMSS scale scores and scale scores estimated for the paper bridge were examined. Similar to the results from the percent correct analyses, country differences in average scale scores between eTIMSS and the paper bridge were small, and few were flagged as statistically significant. Because the bridge scale scores were based on smaller samples and fewer items than the eTIMSS scale scores, these differences may be due to sample differences in addition to residual differences that were present in the linked scales.

This chapter is intended to encourage researchers interested in examining how mode effects can differ among countries, types of items, or student groups. More in depth studies by country experts may be worthwhile to explore to what extent differences in performance between paper bridge and eTIMSS can be attributed to residual mode effects versus sample differences.

References

- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1), 289–300.
- Box, G. E. P., Hunter, J. S., & Hunter, W.G. (2005). *Statistics for experimenters: Design, innovation, and discovery, Second Edition*. Hoboken, New Jersey: John Wiley & Sons, Inc.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7, 1–26. doi:10.1214/aos/1176344552
- Efron, B. (2011) Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106, 1602–1614.
- Fishbein, B., Martin, M. O., Mullis, I. V. S., & Foy, P. (2018). The TIMSS 2019 item equivalence study: Examining mode effects for computer-based assessment and implications for measuring trends. *Large-scale Assessments in Education*, 6(1), 11. <https://doi.org/10.1186/s40536-018-0064-z>
- Shaffer, J. P. (1995). Multiple hypothesis testing. *Annual Review of Psychology*, 46, 561–584. doi:10.1146/annurev.ps.46.020195.003021
- von Davier, M., Yamamoto, K., Shin, H-J., Chen, H., Khorramdel, L., Weeks, J., Davis, S., Kong, N., & Kandathil, M. (2019a). Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assessment in Education: Principles, Policy & Practice*, 26(4), 466–488. <https://doi.org/10.1080/0969594X.2019.1586642>
- von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2019b). Developments in psychometric population models for technology-based large-scale assessments: An overview of challenges and opportunities. *Journal of Educational and Behavioral Statistics*, 44(6), 671–705. <https://doi.org/10.3102/1076998619881789>

Appendix 13A

Comparing Country Level Proportion Correct to International Averages

Consider the international average of a statistic, for example an average proportion correct over a number of item responses. In our case, these are the items that were designed for a paper based assessment, TIMSS 2015, and that were still used in TIMSS 2019 as trend items for computer based countries, and were re-implemented for computer delivery for countries that chose to use the eTIMSS assessment. These trend items were carefully designed for computer delivery so that a majority of 80% or more per grade and subject domain was considered equivalent in terms of how they relate to achievement on the TIMSS scale.

The international average of the average percent-correct typically based on equal contribution of all participating countries, that is, they are defined as an unweighted average. Formally, we have

$$\mu_I = \frac{1}{C} \sum_{c=1}^C \mu_c.$$

Obviously, we do not have the true population values at the country level, as we only collect a sample of schools, and 1 or 2 classrooms per school. The best estimate of the country average percentages are the weighted estimates of the proportion correct, i.e., the weighted sum of correct responses, divided by the sum of weights, over the items that are considered comparable.

The international estimate \widehat{M}_I of this proportion correct has estimation error as well, as it is also based on sampling, albeit over multiple countries. We denote the standard error associated with this average by \widehat{S}_I . Assuming unbiased sample-based estimates, we have

$$E(\widehat{M}_I) = E\left(\sum_{c=1}^C \widehat{M}_c\right) = \mu_I$$

with estimates of country means \widehat{M}_c that are based on the country sample, we also assume these are unbiased, i.e.,

$$E(\widehat{M}_c) = \mu_c$$

and denote the associated standard errors by \widehat{S}_c . For an estimate of the difference, $\widehat{d}_k = \widehat{M}_k - \widehat{M}_I$ of a country k 's mean and overall mean $\Delta_k = \mu_k - \mu_I$ we observe the following complication. The estimate of the international mean \widehat{M}_I contains the country mean \widehat{M}_k as one component. This implies

$$\widehat{S}_{d(k)} = \sqrt{\widehat{S}_I^2 + \widehat{S}_k^2 - 2cov(\widehat{M}_I, \widehat{M}_k)}$$

with

$$\text{cov}(\widehat{M}_I, \widehat{M}_k) = \text{cov}\left(\frac{1}{C} \sum_{c=1}^C \widehat{M}_c, \widehat{M}_k\right) = \frac{1}{C} \text{cov}(\widehat{M}_k, \widehat{M}_k) = \frac{1}{C} \widehat{S}_k^2.$$

Plugging this result into the estimate provides

$$\widehat{S}_{d(k)} = \sqrt{\widehat{S}_I^2 + \widehat{S}_k^2 - \frac{2}{C} \widehat{S}_k^2} = \sqrt{\widehat{S}_I^2 + \left[\frac{C-2}{C}\right] \widehat{S}_k^2}$$

which is well defined whenever there are at least two countries, i.e., whenever $C \geq 2$.

Country Mode Differences, Corrected for International Mode Differences

The international estimate and the expected values of proportion correct of paper items ('P'-samples) will be denoted by

$$E(\widehat{M}_{IP}) = \mu_{IP}$$

and the mean of proportion correct across computer based ('E'-samples) is

$$E(\widehat{M}_{IE}) = \mu_{IE}.$$

Similarly, we have associated standard errors for the estimate of the international proportion correct for paper, S_{IP} , and computer, S_{IE} , respectively, as we have for the country level estimates S_{kP} and S_{kE} . These can be calculated separately using the jackknife procedures and defined as given above. The bridge and the eTIMSS samples do provide an estimate \widehat{d}_{P-E} of the mode difference

$$\Delta_{P-E} = \mu_{IP} - \mu_{IE}$$

at the international level. This mode difference is being controlled for in the linking design that uses the bridge and eTIMSS samples in a customary equivalent groups approach. That means this difference is no longer relevant and can be taken out of country level comparisons of the effect of mode on achievement results. Only any remaining differences that are based on differences at the country levels are relevant, as the overall difference is no longer affecting the plausible values that are provided in the international database.

That means, in order to examine whether there is a difference between the paper- and the eTIMSS proportion correct at the country-level that goes beyond what would be expected internationally. Only differences that go beyond this are relevant, as the international average of correct response differences is already taken care of by the mode effect adjustment. Consequently, the difference

$$\widehat{d}_{Pk-Ek} = \widehat{d}_{Pk} - \widehat{d}_{Ek} \approx \widehat{M}_{Pk} - \widehat{M}_{Ek} - \Delta_{P-E}$$

quantifies the relative paper versus eTIMSS difference of proportions correct that not accounted for by the international linking in the bridge study. For this estimated difference, we can use the standard error

$$\hat{S}_{Pk-Ek} = \sqrt{\hat{S}_{Pd(k)} + \hat{S}_{Ed(k)}}$$

With the estimates defined as above

$$\hat{S}_{Pd(k)} = \sqrt{\hat{S}_{IP}^2 + \left[\frac{C-2}{C}\right] \hat{S}_{kP}^2}$$

and

$$\hat{S}_{Ed(k)} = \sqrt{\hat{S}_{IE}^2 + \left[\frac{C-2}{C}\right] \hat{S}_{kE}^2}$$

Note that these are almost the same as the s.e. for the country mean proportions correct for paper versus eTIMSS, calculated separately. This statistic is adjusted by the s.e. for the international proportion correct (separately calculated by mode) but adjusted for the number of countries included in the international mean proportions.

Achievement data comparisons based on Bridge and eTIMSS samples

The comparison, once the linking is accomplished, is rather straightforward. The standard error estimates for the bridge sample averages and the eTIMSS averages can be used to calculate the standard error of the difference for countries where schools were selected to test either using the paperTIMSS or the eTIMSS assessment. These can, within countries, be assumed to be independent samples, and if the schools were randomly assigned to the mode of assessment, these independent samples can be assumed to be identically distributed. In practice, this may not be completely true, as schools that were able to test on computer may have been somewhat different from schools that were assigned to test with the bridge/paperTIMSS instruments. For example, hardware availability in schools may be associated with average socio-economic status of students' parents.

Assuming independent samples from the same population, the mean difference

$$\mu_{cB} - \mu_{cE} = \Delta$$

Between bridge sample (B) and eTIMSS sample (E) in country c can be evaluated using the standard error of the difference for independent samples,

$$s.e. (\Delta) = \sqrt{s.e. (\mu_{cB})^2 + s.e. (\mu_{cE})^2}$$

However, this is no longer appropriate and may overestimate the s.e. if students were assigned to paper or eTIMSS within schools. In this case, samples are dependent, and the difference of the achievement per school needs to be calculated and the variance of this difference needs to be estimated using an appropriate resampling method (Efron, 1979). The bridge and the eTIMSS samples would in some countries be drawn in the same schools, but different classes, while in other countries the two samples would come from schools without overlap, while a third set of countries would have some schools that assign one class to paperTIMSS and another to eTIMSS, and other schools would only assign one class to one of the modes. The assumption of independent samples is applicable in the case that the different classes perform independently of being sampled in the same or in different schools. If schools are very different compared to between class differences within schools (i.e., there is tracking between schools, but little tracking within schools) this will lead to overestimation of standard errors.

In the exhibits, we assume for simplicity of exposition, we assume independent samples of students taking the eTIMSS and the paperTIMSS assessment.