CHAPTER 12

Examining Country-Level Differences Between digitalPIRLS Data and Bridge Data

Liqun Yin
Bethany Fishbein
Ummugul Bezirhan
Pierre Foy
Matthias von Davier

Introduction

PIRLS 2021 marked the transition from paper-based assessment to digital-based assessment, with about half of the countries choosing to administer the digital format (digitalPIRLS) and the other half remaining on paper (paperPIRLS). The TIMSS & PIRLS International Study Center made every effort to ensure a seamless transition from the PIRLS paper-based format to the digital-based format, beginning with instrument development (see Chapter 1). In addition to administering the digital assessment to the full PIRLS samples of students, digitalPIRLS countries also administered paper booklets of "trend" texts and items to a smaller equivalent "bridge" sample to link the digital data to the PIRLS trend scale. The purpose of this chapter is to guide digitalPIRLS countries and other secondary users of the data in comparing the digitalPIRLS data and the paper-based bridge data to examine how the linking adjustment may have affected the PIRLS 2021 achievement results.

As described in PIRLS 2021 Assessment Design, digitalPIRLS adopted the same booklet design as paperPIRLS, with the same 18 text and item sets assembled into 18 student booklets so that each contained one "literary" text and one "informational" text. digitalPIRLS also integrated five ePIRLS online informational reading tasks, requiring additional booklets—20 booklets with two ePIRLS tasks and 45 booklets with one ePIRLS task and one paper-equivalent informational text. To collect data from the bridge samples, eight trend text and item sets that were administered in PIRLS 2016 were assembled into eight paper booklets.

The bridge samples that received the trend texts in the paperPIRLS assessment are randomly equivalent to their full digitalPIRLS sample counterparts, having been drawn from the same student populations. It will be shown in this chapter that the digitalPIRLS and bridge samples are very similar when compared on a number of key indicator variables. As such, the bridge data form a

https://doi.org/10.6017/lse.tpisc.tr2103.kb9281

link between computer-based data in 2021 and the data of paperPIRLS countries in 2021, as well as paper-based data in 2016.

While the main purpose of the bridge sample was to facilitate equivalent groups linking between the paperPIRLS and digitalPIRLS achievement data at the international level, the bridge data also allow for examining differences between paper-based data and digital-based data at the country level. However, it is important to note that the bridge sample received only trend texts, and is only about one-third of the size of the digitalPIRLS sample. Therefore, results of country-level analyses need to be understood as tendencies rather than definitive indicators of mode differences, which would require a much larger bridge sample size and would not allow for integrating ePIRLS into the assessment.

The PIRLS 2021 Bridge Between digitalPIRLS and paperPIRLS

In PIRLS 2021, all countries transitioning to digitalPIRLS included a bridge sample that facilitated linking to paperPIRLS. The eight bridge booklets were administered to an additional sample of 1,500 students, randomly sampled from the same population, and in many instances from the same schools, as the full digitalPIRLS sample. The same administration procedures and testing conditions were applied to the two randomly selected samples, including the administration of the same set of contextual questionnaires.

As described in Chapter 10, using the bridge data as the link between paperPIRLS and digitalPIRLS is an example of equivalent groups design, a well-researched and frequently applied linking approach that is commonly referred to as a "randomly equivalent samples" design (Dorans & Puhan, 2017; Haberman, 2015; Kolen & Brennan, 2014; M. von Davier & A. von Davier, 2007). The TIMSS & PIRLS International Study Center has previously used the equivalent samples linking approach, such as when linking a PIRLS colorized Reader in early PIRLS assessments (Gonzalez, 2003) and when changing the TIMSS assessment design in 2007 (Foy et al., 2008). Because PIRLS 2021 students were randomly selected for the digitalPIRLS sample or bridge sample from the same student population and in the same manner, the students taking these assessment formats can be expected to have the same distribution of underlying skills and knowledge, with only small differences due to sampling. They are otherwise equivalent, differing only in that they were randomly assigned to different testing modes.

Under the equivalent groups design, having a substantial percentage of the same PIRLS assessment blocks in both paper and digital modes strengthens the validity and interpretability of achievement results based on linking the two test modes. As described in Chapter 1, the TIMSS & PIRLS International Study Center made significant efforts in developing the user interface for digitalPIRLS to ensure it was easy and intuitive for students to navigate between text screens and

between items. The interface included a highlighter tool for students to highlight parts of the text, similarly to how students mark or underline parts of text on paper.

Country-Level Differences in Average Percent Correct on PIRLS 2021 Trend Items between digitalPIRLS and Bridge Samples

When using the random equivalent groups design, a difference between group-level performance is taken as an indication of a difference in difficulty between the two test formats (Kolen & Brennan, 2014). An observed difference in reading performance based on the trend items between the bridge and digital samples at the international level indicates a difference in level of difficulty between the two testing modes. This means that, as long as the equivalence between the two randomly selected groups is valid and while accounting for sampling error, it can be inferred that students' responses to the trend items were impacted by the change of administration mode.

To help users of the PIRLS 2021 data gain an understanding of the differences observed when moving from paperPIRLS to digitalPIRLS, the analyses in this section compare the average performance between the paper bridge and digital data based on the 117 trend items that were common between modes. Apart from random sampling differences and minor deviations from the sampling design that might have caused some departure from this equivalence of comparison groups at the country level, the average performance differences between the bridge and digitalPIRLS data can be attributed to a systematic difference between the two modes of administration at the international level. Because PIRLS is an international study, this international difference is central and needs to be considered when examining country-level differences and adjusted for through the equivalent groups design in linking. The approach presented here provides a model for investigating country mode differences for different types of items or student groups. The relatively straightforward computations used here are described in Appendix 12A.

Exhibit 12.1 shows each digital country's average performance on the trend items for the paper bridge and digital samples as well as the average across countries. The average in the first panel of Exhibit 12.1 is based on the 18 digital countries that assessed the fourth grade students at the end of the school year.[1] The average in the second panel is the average across digital countries that delayed test administrations and assessed the fourth grade cohort at the beginning of fifth grade.[2]

1 Because of the COVID-19 disruption, not all participating countries managed to administer the PIRLS assessment at the target fourth grade cohort at the scheduled time. Only the digitalPIRLS trend countries which administered the PIRLS 2021 assessment at the end of fourth grade and according to the original scheduled time were included in item calibration and equivalent groups linking models.

2 The United States administered the PIRLS 2021 digital assessment and the PIRLS 2021 paper bridge assessment. The United States opted to report the paper bridge results since, as a delayed testing country, their data were treated as non-trend when conducting the population-based linking between the paper and digital data.

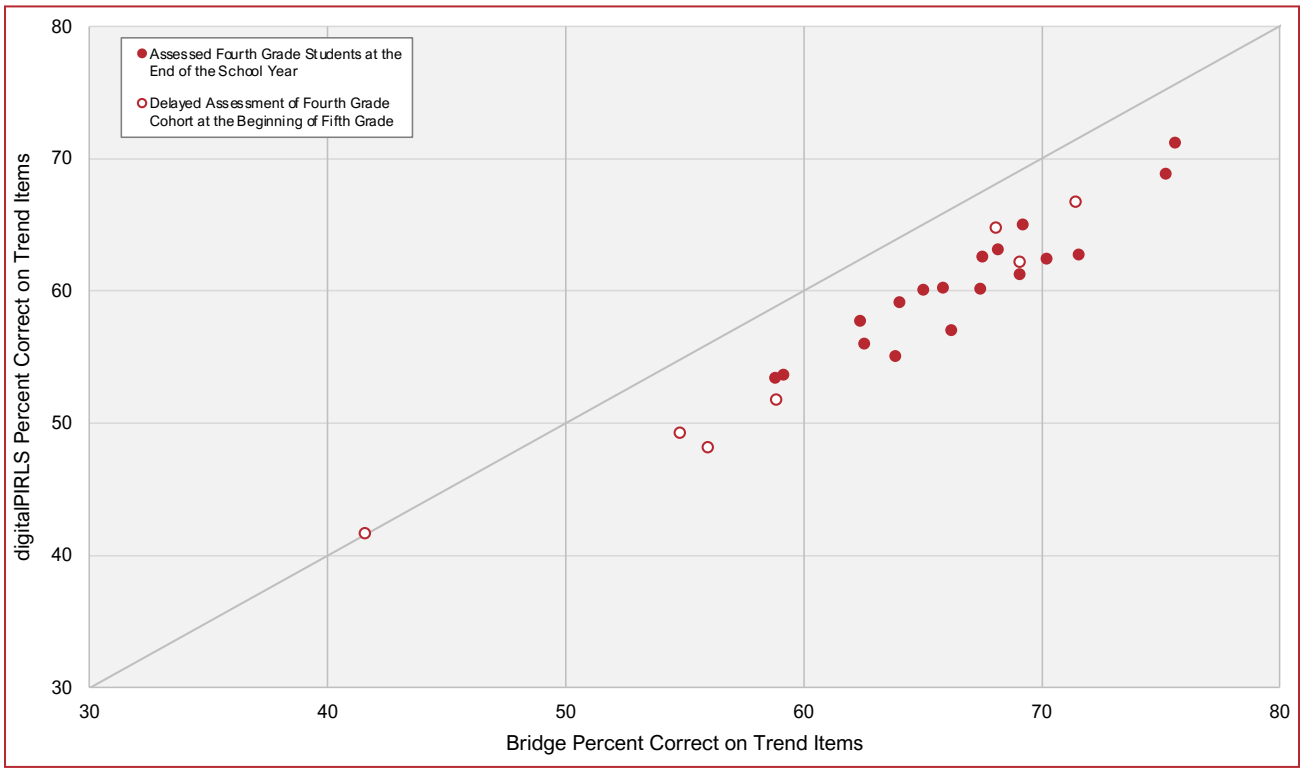
Exhibit 12.1: Average Percent Correct Across Trend Items for Digital and Bridge Samples

Country	Bridge Sample	digitalPIRLS Sample
Assessed Fourth Grade Students at the End of the School Year		
Belgium (Flemish)	59.14 (0.89)	53.67 (0.64)
Chinese Taipei	71.56 (0.64)	62.78 (0.53)
Czech Republic	70.19 (0.65)	62.48 (0.71)
Denmark	68.16 (0.69)	63.12 (0.59)
Finland	69.21 (0.88)	65.00 (0.61)
Germany	67.41 (0.79)	60.18 (0.59)
Israel	63.84 (0.82)	55.12 (0.65)
Italy	69.08 (0.68)	61.29 (0.58)
Malta	58.79 (1.49)	53.42 (0.73)
New Zealand	62.35 (0.84)	57.76 (0.69)
Norway (5)	65.00 (0.77)	60.10 (0.64)
Portugal	66.19 (0.78)	57.06 (0.59)
Russian Federation	75.22 (0.93)	68.88 (0.91)
Singapore	75.60 (0.90)	71.24 (0.74)
Slovak Republic	65.86 (1.08)	60.24 (0.75)
Slovenia	64.01 (0.77)	59.13 (0.56)
Spain	62.51 (0.76)	56.03 (0.64)
Sweden	67.50 (0.89)	62.60 (0.61)
Average (18)	**66.76 (0.20)**	**60.56 (0.16)**
Delayed Assessment of Fourth Grade Cohort at the Beginning of Fifth Grade		
Croatia	71.41 (0.98)	66.79 (0.83)
Hungary	69.05 (1.05)	62.21 (0.82)
Kazakhstan	58.83 (0.89)	51.79 (0.68)
Lithuania	68.05 (0.91)	64.81 (0.63)
Qatar	55.98 (1.57)	48.19 (0.82)
Saudi Arabia	41.55 (1.45)	41.71 (0.84)
United Arab Emirates	54.80 (1.65)	49.26 (0.37)
Average (7)	**59.95 (0.47)**	**54.97 (0.28)**

( ) Standard errors appear in parentheses.

The scatterplot in Exhibit 12.2 shows the national average percent correct based on all trend items for the paper bridge and digital samples. The horizontal axis represents the average percent correct across the bridge items. The vertical axis represents the average percent correct of the digital trend items.

Exhibit 12.2: Plot of PIRLS 2021 Country Average Percent Correct Across Trend Items for Digital and Bridge Samples



Exhibits 12.1 and 12.2 show that all (but one) digitalPIRLS countries had higher percent correct on the paper bridge items than on the digital trend items, by about 6 percentage points across countries, on average. Overall, answering items about the PIRLS texts on paper was easier than on the computer. This international difference between average percent correct statistics on the same set of items presented in the two modes revealed a non-negligible difference between the paper and digital assessments.

These findings generally are consistent with other studies about assessments in reading (e.g., Clinton, 2019; Kong et al., 2018), as well as in mathematics and science as shown in TIMSS (Fishbein et al., 2018; von Davier et al., 2020). As discussed by Fishbein et al. (2018), Jerrim et al. (2018), and other researchers, there are several potential causes of this mode difference. They include differences between reading on paper versus on a screen, differences in test-taking strategies, advantages or disadvantages for students in locating information and typing responses,

technical difficulties with computer administration of tests in schools, and differences in student engagement during the test session. Among other factors, they are potential contributors to a mode-related performance difference. To ensure country achievement distributions can be properly evaluated and compared, this requires a linking methodology that can adequately account for this difference at the international level.

Exhibits 12.3 and 12.4 enable a comparison of the bridge and digital samples across countries that does not confound the international average effect with country-level mode differences by showing the country-level bridge-digital differences adjusted for the international average percent correct difference (see Appendix 12A). The results are presented in tabular form and graphical form, respectively. In these exhibits, the averages based on the top panel of Exhibit 12.1 (66.76% for bridge, 60.56% for digital) were used as the international baseline to adjust all countries, including countries with delayed assessment where students were half a year older (see Chapter 8).

Exhibit 12.3: Deviations from International Baseline Average Percent Correct Across Trend Items for Digital and Bridge Samples

Country	Bridge Sample	digitalPIRLS Sample	Difference	
Assessed Fourth Grade Students at the End of the School Year				
Belgium (Flemish)	-7.62 (0.86)	-6.89 (0.62)	0.73 (1.07)	
Chinese Taipei	4.80 (0.63)	2.22 (0.52)	-2.58 (0.82)	▼
Czech Republic	3.43 (0.65)	1.92 (0.69)	-1.52 (0.94)	
Denmark	1.40 (0.68)	2.56 (0.58)	1.16 (0.89)	
Finland	2.46 (0.86)	4.44 (0.59)	1.98 (1.04)	
Germany	0.65 (0.77)	-0.38 (0.58)	-1.04 (0.96)	
Israel	-2.91 (0.80)	-5.44 (0.63)	-2.53 (1.02)	▼
Italy	2.32 (0.68)	0.73 (0.56)	-1.59 (0.88)	
Malta	-7.97 (1.42)	-7.14 (0.70)	0.83 (1.58)	
New Zealand	-4.40 (0.82)	-2.80 (0.67)	1.61 (1.06)	
Norway (5)	-1.76 (0.75)	-0.46 (0.62)	1.30 (0.98)	
Portugal	-0.57 (0.76)	-3.50 (0.58)	-2.93 (0.96)	▼
Russian Federation	8.46 (0.90)	8.32 (0.87)	-0.15 (1.25)	
Singapore	8.85 (0.88)	10.68 (0.71)	1.84 (1.13)	
Slovak Republic	-0.90 (1.04)	-0.32 (0.72)	0.58 (1.27)	
Slovenia	-2.74 (0.76)	-1.43 (0.55)	1.31 (0.94)	
Spain	-4.24 (0.75)	-4.53 (0.62)	-0.28 (0.97)	
Sweden	0.74 (0.86)	2.04 (0.59)	1.30 (1.05)	
Average (18)	**0.00**	**0.00**	**0.00**	

Exhibit 12.3: Deviations from International Baseline Average Percent Correct Across Trend Items for Digital and Bridge Samples (Continued)

Country	Bridge Sample	digitalPIRLS Sample	Difference	
Delayed Assessment of Fourth Grade Cohort at the Beginning of Fifth Grade				
Croatia	4.66 (1.00)	6.23 (0.85)	1.57 (1.31)	
Hungary	2.29 (1.07)	1.65 (0.84)	-0.64 (1.36)	
Kazakhstan	-7.92 (0.92)	-8.77 (0.70)	-0.85 (1.15)	
Lithuania	1.29 (0.93)	4.25 (0.65)	2.96 (1.14)	▲
Qatar	-10.77 (1.58)	-12.37 (0.84)	-1.60 (1.79)	
Saudi Arabia	-25.21 (1.46)	-18.85 (0.85)	6.36 (1.69)	▲
United Arab Emirates	-11.96 (1.66)	-11.31 (0.40)	0.65 (1.71)	
Average (7)	**-6.80**	**-5.60**	**1.21**	

( ) Standard errors appear in parentheses. Because of rounding some results may appear inconsistent.
▲ indicates digitalPIRLS performance significantly higher than bridge (α = 0.05)
▼ indicates digitalPIRLS performance significantly lower than bridge (α = 0.05)

Exhibit 12.4: Plot of Country Deviations from International Baseline Average Percent Correct Across Trend Items for Digital and Bridge Samples

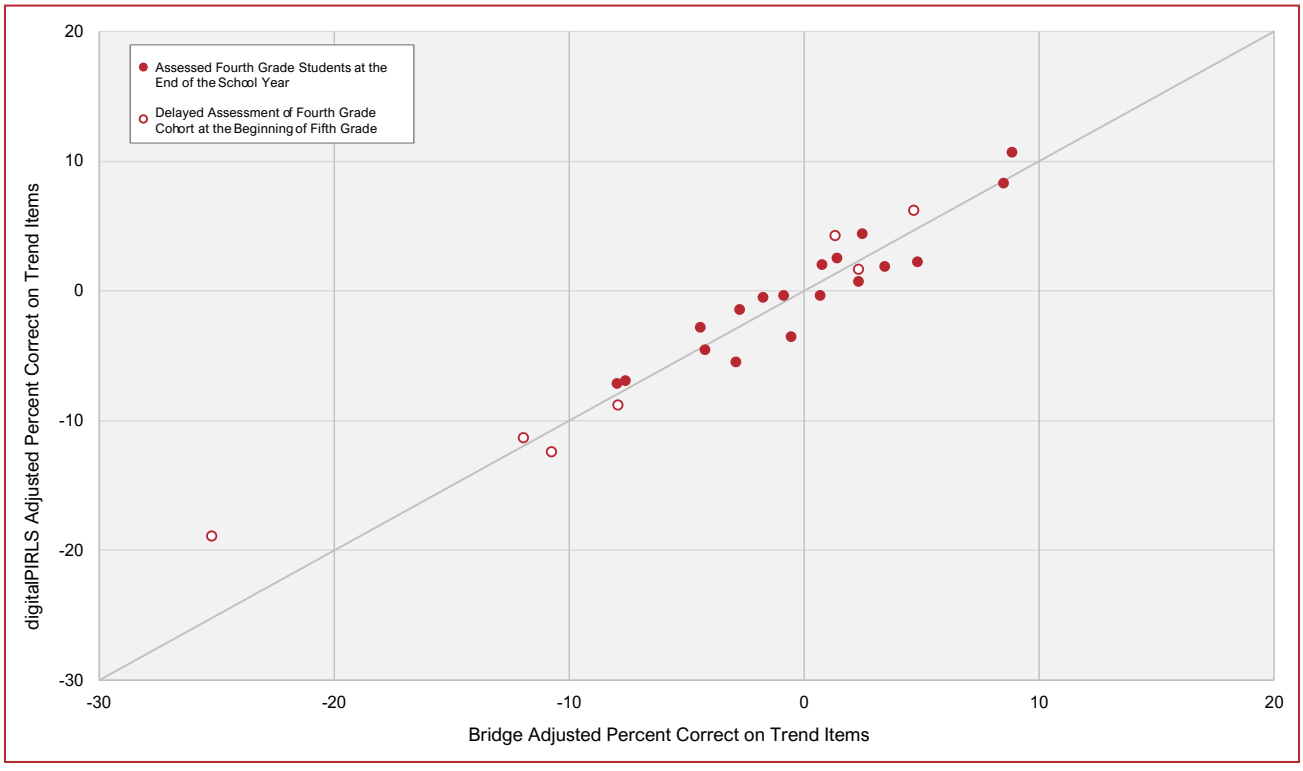


Exhibit 12.3 shows the country deviations from the international baseline for the paper bridge and digital samples, together with their standard errors. The bridge column shows the difference from the bridge baseline, the average bridge proportion correct (66.76%), and the digital column shows the difference from the digital baseline average (60.56%). The third column can be viewed as a variation of the difference in differences (DD) method, which allows an evaluation of whether the bridge and digital samples differ more than expected based on the international baseline (described in Appendix 12A). For example, Belgium's (Flemish) deviation for the bridge was –7.62 (0.86), and for digital was –6.89 (0.62). The relative difference for the country is the difference between the two deviations, e.g., 0.73 for Belgium (Flemish), which is not significant. The relative difference provides an estimate of the country mode difference adjusted for the international baseline difference between modes.

Exhibits 12.3 and 12.4 provide a way to evaluate whether each country had positive or negative bridge-digital difference beyond the international difference. Although most differences were not statistically significant given their standard errors, there were some significant differences ($p < 0.05$), mostly small, after adjusting for the international average difference. Among the five countries with statistically significant differences, three in the top panel performed better in the paper bridge than in digitalPIRLS, relative to the average international difference. The other two in the bottom panel performed better in digital. Note that among those countries that performed relatively better in digitalPIRLS, all tested older students in a delayed window, at the beginning of fifth grade, due to COVID-19 disruptions.

Rationale for Population-Based Linking of Paper and Digital Assessments

This section provides an overview and rationale for the population-based linking utilized in PIRLS 2021 to adjust for the difference between forms by relying on the randomly equivalent samples. First, after this overview, the differences between paperPIRLS and digitalPIRLS are evaluated at the item level using item response theory (IRT) parameters. Then, analyses examine the equivalence of bridge and digital samples based on sampling outcomes and key context variables. The section concludes with a summary of the population-based linking adopted in PIRLS 2021 to link the paper-based and computer-based assessments. Chapter 10 provides more detailed descriptions of the methodology, and Chapter 11 describes the implementation procedures.

Both TIMSS 2019 and PIRLS 2021 implemented a data collection design meant to ensure two of the most frequently used approaches to IRT-based linking are feasible: item-based linking as well as equivalent samples linking. The item-level linking approach relies on "mode effect modeling" (von Davier et al., 2019), assuming measurement invariance across items administered in both modes. Item-level linking makes a relatively strong assumption that a substantive number

of invariant items between modes can be identified and used to calculate an international mode adjustment for item difficulty to account for achievement differences between administration modes.

Item-based linking in TIMSS 2019 was a success and required accounting for only small differences in item difficulties between modes for a large number of invariant items that equated to a magnitude of 5 to 10 points on the TIMSS trend scales (Foy et al., 2020). This required a substantial number of items to retain invariance properties as described in the research literature on measurement invariance (e.g., Millsap, 2011; von Davier, 2020). A prerequisite for the validity of this item-based linking approach is the presence of a large set of items constructed to be invariant by design between modes based on item content, format, and expected psychometric functioning.

However, based on psychometric analyses of the PIRLS 2021 data, the differences between PIRLS 2021 paper and digital item-level data were much less homogenous than those in TIMSS 2019. The item equivalence assumption did not hold in PIRLS 2021, which showed a wider variety of differences in how items functioned between modes. The subsections below show examples of item characteristic functions illustrating how some items were easier in digital format, some were about the same difficulty in digital and paper formats, but most items were more difficult in the digital format.

Therefore, PIRLS 2021 relied on the randomly equivalent samples approach to link the paper and digital assessments. This population-based linking did not require assuming item invariance and was supported by the PIRLS 2021 data collection design, which included representative random samples from the same populations taking either the paper-based or the computer-based versions of the assessment. Underlying this approach is the principle of randomization, one of the central building blocks of experimental design (Box et al., 2005), which aims to ensure that observed differences in results of groups exposed to different treatments are due to the treatment differences and not pre-existing differences between the groups.

The following sections provide details and rationale to justify the population-based linking approach adopted in PIRLS 2021.

Evaluating Item-Level Differences

To evaluate differences in item functioning between paperPIRLS and digitalPIRLS, the TIMSS & PIRLS International Study Center used multiple-group item response theory (MG-IRT; Bock & Zimowski, 1997) to estimate separate sets of paper and digital item parameters and compare them. This psychometric model combined digital data with all available paper data, including PIRLS 2016, PIRLS 2021, and bridge data, in a model that estimated all item parameters freely for both paper and digital items on the same metric. The PIRLS 2016 and 2021 calibration countries served as distinct groups (allowing population differences in achievement) in the MG-IRT model, while the bridge and digital samples from 2021 were considered equivalent samples from the same

groups. Applying this model put the digitalPIRLS items on the same scale as the paper assessment items, relying on the randomly equivalent bridge and digital samples drawn from the same target populations, while treating digital items as different from paper items. Chapter 11 provides a more detailed description of this analysis.

A comparison of paper and digital item parameters reveals the differences in item functioning between modes in terms of difficulty (location), as shown in Exhibit 12.5; discrimination (slope) as shown in Exhibit 12.6; and guessing, as shown in Exhibit 12.7. In each exhibit, the horizontal axis represents the IRT-based item parameters for paper items, and the vertical axis represents the item parameters for the corresponding digital items. In the presence of invariant items, item difficulty parameters are expected to cluster along a line parallel to the reference diagonal in Exhibit 12.5. Also, item discrimination and guessing parameters are expected to be similar, clustering along the reference diagonal in Exhibits 12.6 and 12.7, respectively. Instead, all three exhibits show large variations in item parameters between modes.

Exhibit 12.5: Plot Comparing Item Difficulty Parameters Between Paper and Digital Administrations

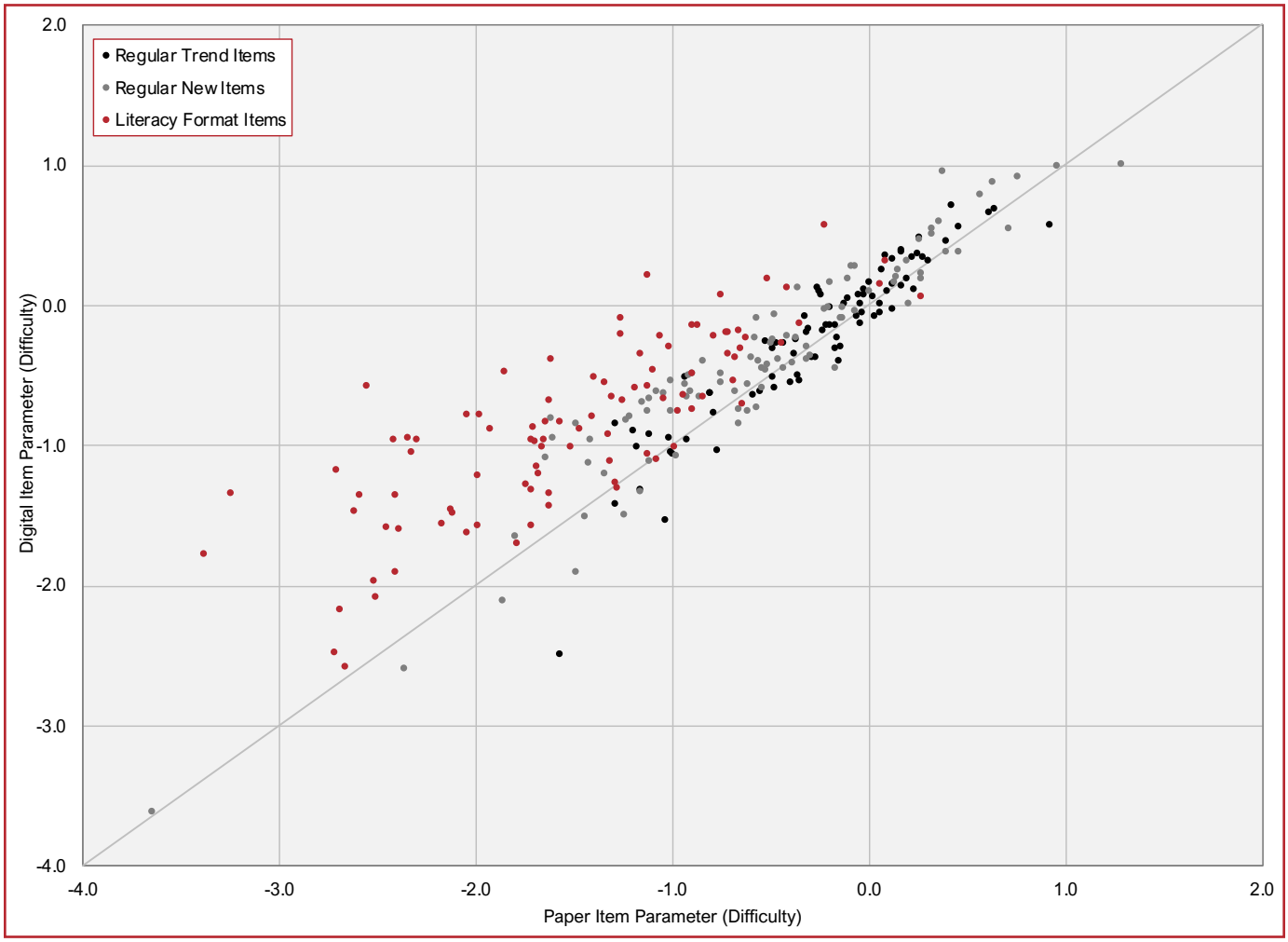


Exhibit 12.6: Plot Comparing Item Discrimination Parameters Between Paper and Digital Administrations

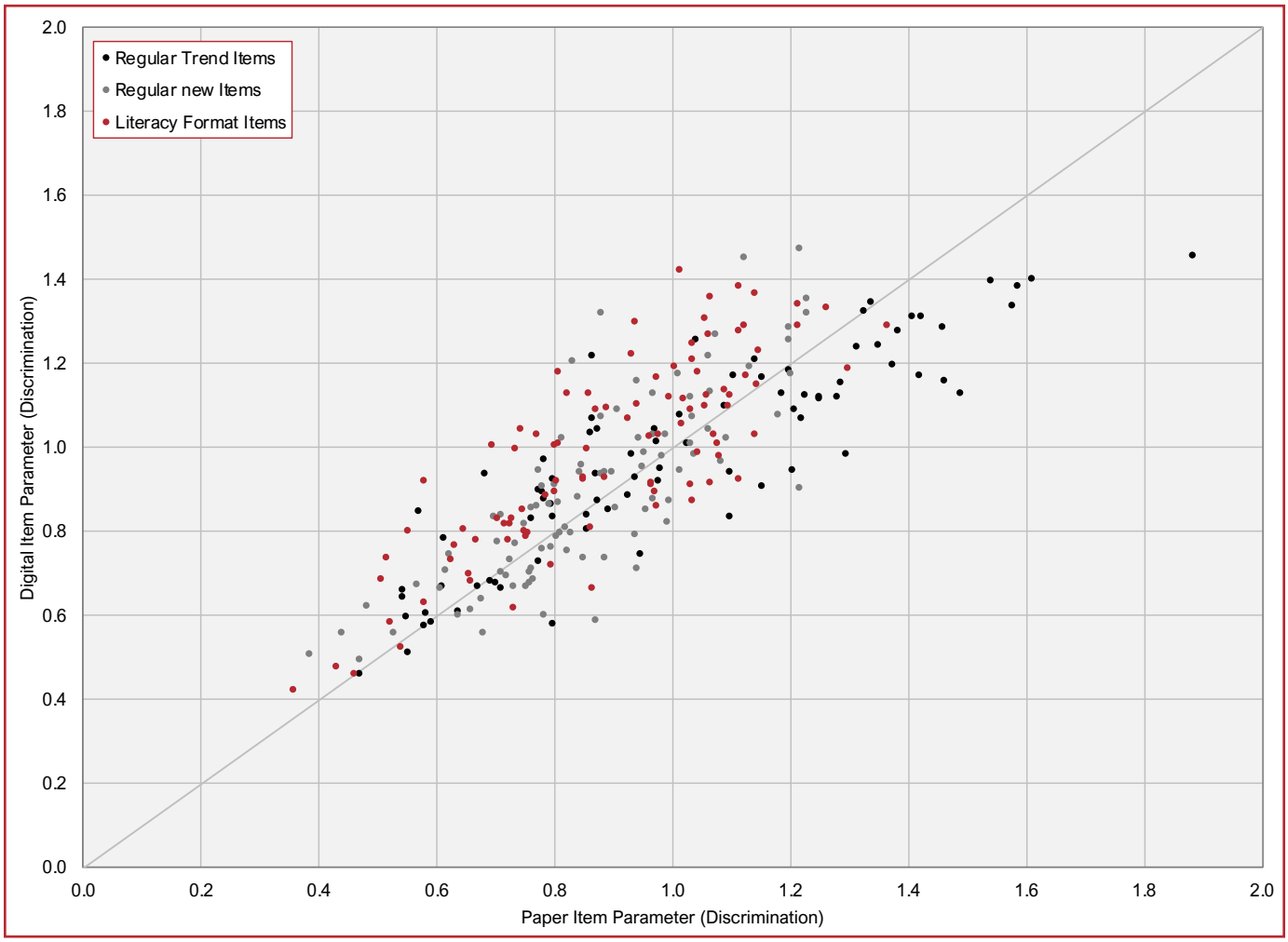
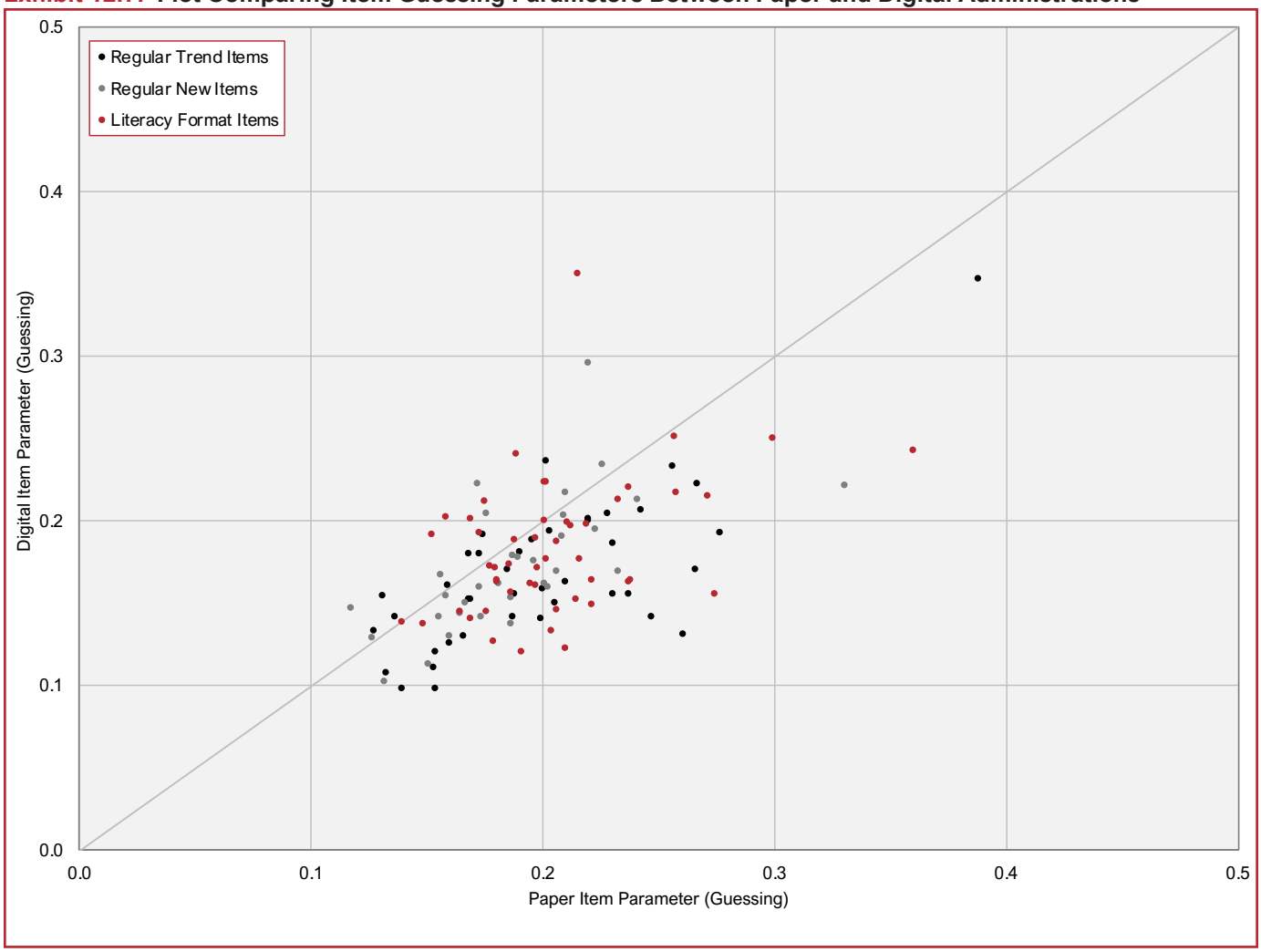


Exhibit 12.7: Plot Comparing Item Guessing Parameters Between Paper and Digital Administrations



While many items appeared to be easier in paper format (data points above the reference diagonal in Exhibit 12.5), some items were easier in digital format (data points below the reference diagonal), and a few items showed no apparent difference (items at or near the reference diagonal). With a correlation coefficient of 0.88 between paper and digital difficulty parameters, there was no clear pattern, let alone a uniform shift, between the two modes. By contrast, the correlation coefficient between paper and digital difficulty parameters for TIMSS 2019 was 0.95 for both fourth grade mathematics and fourth grade science, respectively. Moreover, there was little uniformity to be found in the discrimination and guessing parameters in Exhibit 12.6 and 12.7, with correlation coefficients of 0.82 and 0.59, respectively. In TIMSS 2019, these correlation coefficients were 0.94 and 0.82, respectively, for fourth grade mathematics; 0.88 and 0.83, respectively, for fourth grade science.

It is noteworthy that the items included from texts developed for PIRLS Literacy 2016 (the red data points in the plots) showed apparently larger item difficulty differences than the items from the other texts. This may be because PIRLS Literacy item sets underwent format changes when they were adapted for computer delivery. PIRLS Literacy passages were developed to be easier, and the texts and items were split up into portions such that each item corresponded to only a small portion of the full text shown on the opposite page of the open booklet. When the PIRLS Literacy 2016 paper texts and items were converted to digital versions, the presentation had to be changed to harmonize the functionality of PIRLS Literacy and "regular" PIRLS passages for inclusion in the digitalPIRLS assessment. This was required to present test takers with a uniform interface when answering items on the computer, without presenting items directly next to the portion of text to which they pertain. In addition, the passages and items converted from PIRLS Literacy were generally less difficult. The change in item format and mode might have had more impact on the relatively easier PIRLS Literacy items.

The non-uniformity between paper and digital item parameters was also captured by the differences between paper and digital item characteristic curves (ICCs) across items. Exhibits 12.8 and 12.9 show examples of two items with item function differences in opposite directions. Some items exhibited non-uniform differential item functioning. That is, the item functions differed not only in difficulty but also discrimination, so the difference was not the same for students with high and low reading abilities, as shown in Exhibit 12.10. In each plot, the horizontal axis represents the proficiency scale, and the vertical axis represents the probability of a correct response. The fitted curve based on the estimated paper item parameters is shown as a solid red line. The fitted curve based on the estimated digital item parameters is shown as a blue line.

Exhibit 12.8: Example Overlaid Item Characteristic Curves with a Difference Favoring Paper Format

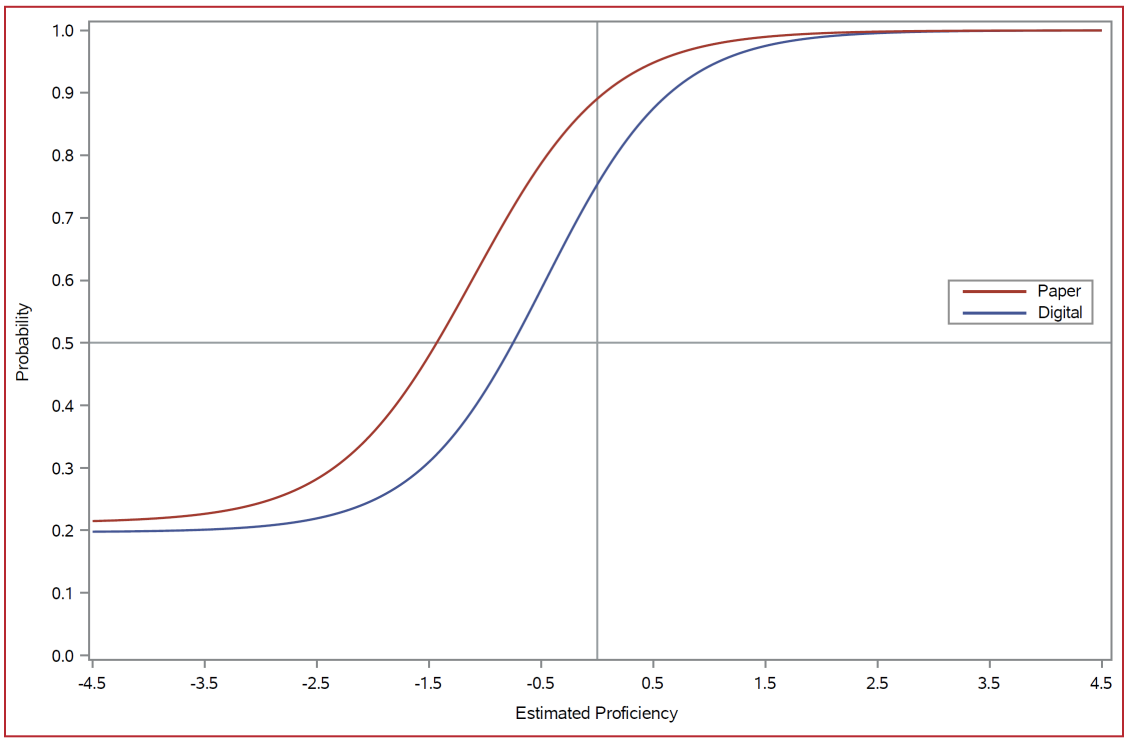


Exhibit 12.9: Example Overlaid Item Characteristic Curves with a Difference Favoring Digital Format

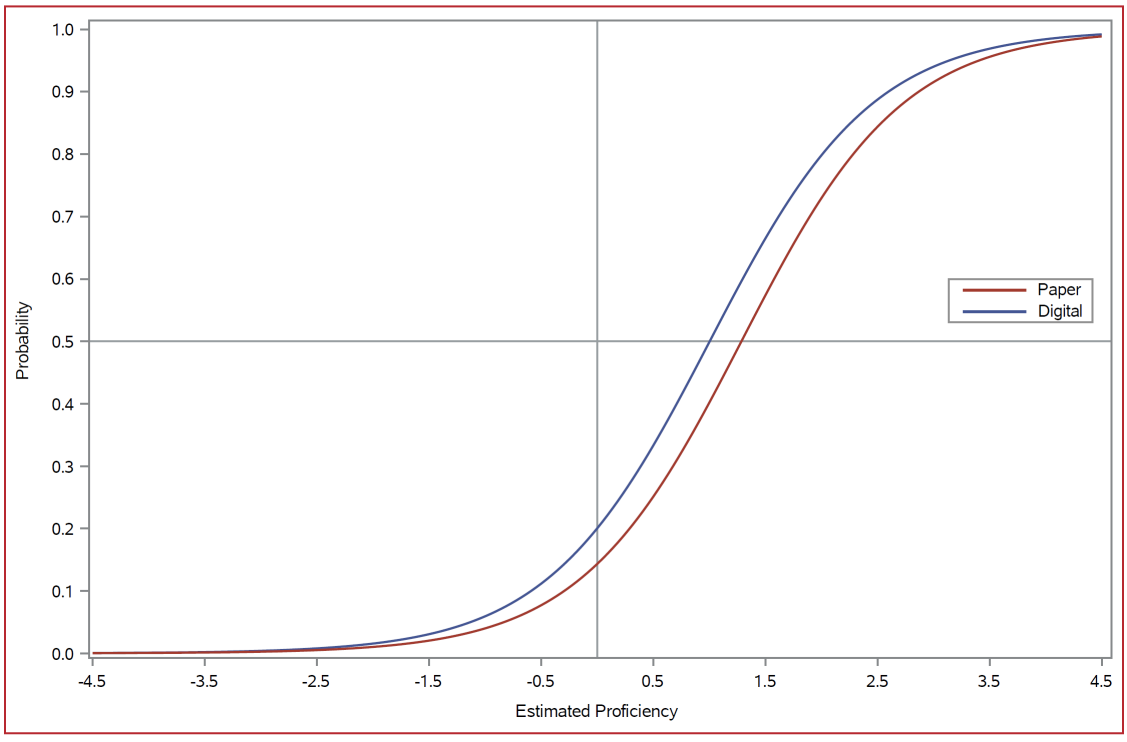
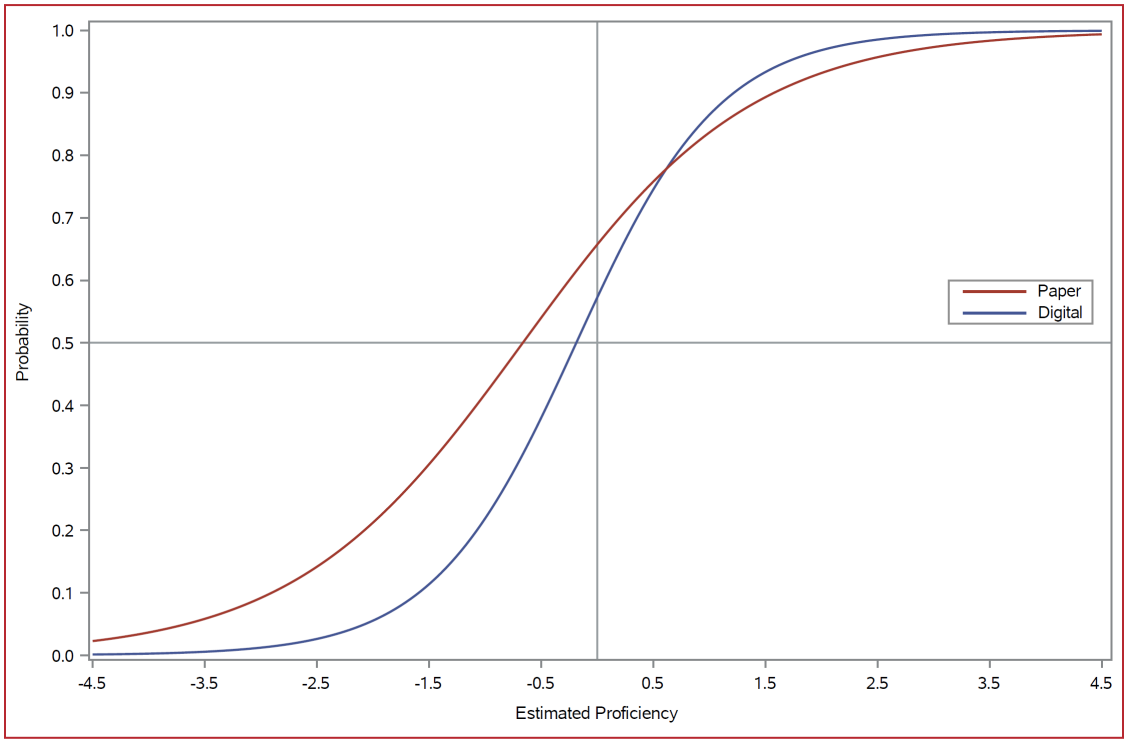


Exhibit 12.10: Example Overlaid Item Characteristic Curves with Non-Uniform Mode Difference between Paper and Digital Formats



Based on the empirical evidence, very few items exhibited similar statistical and psychometric properties between paper and digital formats. Therefore, the item-invariance approach for linking the paper and digital data could not be appropriately applied for linking in PIRLS 2021. The PIRLS 2021 approach also needed flexibility for various factors related to the COVID-19 pandemic (see Chapter 10).

Examining Equivalence of Bridge and Digital Samples

Given the observed differences in average item percent correct and item parameter estimates between paper and digital samples, it is important to establish that the bridge samples and the digital samples, drawn from the same populations, can indeed be considered equivalent samples. Because the bridge and digital samples were randomly selected from the same target population, the two samples are expected to be equivalent in terms of observable and unobservable characteristics. This section presents the results of analyses conducted to evaluate how well the sample characteristics fit the assumption of equivalent samples across digital countries. As is shown below, the evidence collected from comparing sampling outcomes and key background variables between the samples supports the assumption of the random equivalence of the bridge and digital samples.

Sampling Outcomes

To evaluate the comparability of the bridge and digital samples, sampling outcomes were examined for each digital country, as well as across countries. Exhibit 12.11 shows the PIRLS 2021 exclusion and participation rates of the bridge and digital samples. Although there was variation across countries, the differences between the bridge and digital samples were very small.

Exhibit 12.11: Exclusion and Participation Rates of the Digital and Bridge Samples

Country	Overall Exclusion Rates		Participation Rates (Weighted)							
			School (Before Replacement)		School (After Replacement)		Class		Student	
	Digital	Bridge	Digital	Bridge	Digital	Bridge	Digital	Bridge	Digital	Bridge
Assessed Fourth Grade Students at the End of the School Year										
Belgium (Flemish)	2.9%	3.6%	80%	93%	84%	93%	100%	100%	96%	96%
Chinese Taipei	1.1%	1.1%	99%	100%	100%	100%	100%	100%	98%	99%
Czech Republic	5.5%	4.0%	99%	100%	99%	100%	100%	98%	91%	91%
[2] † Denmark	9.1%	8.8%	76%	73%	90%	91%	100%	100%	94%	96%
Finland	2.3%	2.5%	100%	100%	100%	100%	100%	100%	97%	96%
[2] Germany	4.0%	5.8%	95%	98%	97%	98%	100%	100%	88%	88%
[3] Israel	25.7%	24.6%	99%	100%	99%	100%	100%	100%	89%	86%
[2] Italy	5.7%	6.2%	93%	93%	99%	100%	99%	99%	94%	95%
Malta	2.5%	1.3%	100%	100%	100%	100%	100%	100%	90%	88%
† New Zealand	3.5%	3.1%	78%	82%	92%	96%	100%	100%	91%	92%
Norway (5)	4.2%	5.0%	98%	98%	99%	98%	100%	100%	95%	96%
[2] Portugal	6.4%	5.9%	82%	80%	100%	99%	100%	100%	96%	96%
[2] Russian Federation	5.4%	6.7%	99%	100%	100%	100%	100%	100%	97%	97%
[3] Singapore	14.5%	14.5%	100%	100%	100%	100%	100%	98%	97%	98%
† Slovak Republic	2.4%	2.2%	80%	82%	94%	96%	100%	99%	92%	94%
Slovenia	2.8%	2.8%	95%	85%	97%	91%	100%	100%	95%	96%
Spain	4.6%	4.4%	100%	100%	100%	100%	100%	100%	92%	96%
[2] Sweden	5.5%	3.7%	95%	98%	97%	98%	100%	100%	93%	91%
Delayed Assessment of Fourth Grade Cohort at the Beginning of Fifth Grade										
† Croatia	4.4%	3.7%	92%	92%	95%	93%	97%	97%	84%	86%
Hungary	4.9%	5.3%	90%	97%	96%	98%	100%	99%	95%	94%
[2] Kazakhstan	3.9%	8.5%	100%	100%	100%	100%	100%	100%	97%	98%

Exhibit 12.11: Exclusion and Participation Rates of the Digital and Bridge Samples (Continued)

Country	Overall Exclusion Rates		Participation Rates (Weighted)							
			School (Before Replacement)		School (After Replacement)		Class		Student	
	Digital	Bridge	Digital	Bridge	Digital	Bridge	Digital	Bridge	Digital	Bridge
Lithuania	4.5%	3.8%	95%	100%	95%	100%	99%	100%	87%	87%
Qatar	3.1%	3.0%	99%	96%	99%	96%	100%	100%	89%	89%
[3] Saudi Arabia	10.8%	11.3%	95%	96%	100%	100%	100%	100%	93%	95%
United Arab Emirates	4.1%	4.8%	100%	99%	100%	99%	100%	100%	91%	91%
[2] ≡ United States*	7.6%	5.8%	55%	54%	64%	67%	100%	100%	94%	95%

2 National Defined Population covers 90% to 95% of National Target Population.

3 National Defined Population covers less than 90% of National Target Population (but at least 77%).

† Achieved the minimum acceptable participation rates only after including replacement schools.

≡ Did not meet the required sampling participation rates even with the use of replacement schools.

See Chapter 8 for more information about sampling annotations.

* The United States administered the PIRLS 2021 digital assessment and the PIRLS 2021 paper bridge assessment. The United States opted to report the paper bridge results since, as a delayed testing country, their data were treated as non-trend when conducting the population-based linking between the paper and digital data.

Exhibit 12.12 shows the overlap between bridge and digital samples in terms of percentages of students in the bridge sample who were in schools where both digitalPIRLS and bridge booklets were administered. About half of the PIRLS 2021 digital countries had some overlap within schools between their bridge and digital samples. While a sizeable overlap is desirable to strengthen the random equivalence of the bridge and digital samples, it is not necessary, as all national bridge and digital samples were sampled from the same national target populations and under the same sample designs in order to establish their random equivalence (see Chapter 8).

Exhibit 12.12: Percentages of Bridge Samples Overlapping with Digital Samples

Country	Number of Schools	Number of Students	Percentage of Bridge Students in digitalPIRLS Schools (Weighted)
Assessed Fourth Grade Students at the End of the School Year			
Belgium (Flemish)	48	1,623	0.0%
Chinese Taipei	68	1,669	73.4%
Czech Republic	58	1,906	0.0%
Denmark	60	1,403	34.3%
Finland	62	2,069	0.0%
Germany	74	1,343	72.9%

Exhibit 12.12: Percentages of Bridge Samples Overlapping with Digital Samples (Continued)

Country	Number of Schools	Number of Students	Percentage of Bridge Students in digitalPIRLS Schools (Weighted)
Israel	77	1,780	94.6%
Italy	58	1,979	0.0%
Malta	22	835	0.0%
New Zealand	65	2,221	0.0%
Norway (5)	55	1,673	0.0%
Portugal	88	2,098	88.2%
Russian Federation	92	2,187	0.0%
Singapore	60	1,988	100.0%
Slovak Republic	73	1,640	35.4%
Slovenia	51	1,414	34.3%
Spain	74	1,572	53.1%
Sweden	49	1,863	0.0%
Delayed Assessment of Fourth Grade Cohort at the Beginning of Fifth Grade			
Croatia	48	1,226	0.0%
Hungary	52	1,697	0.0%
Kazakhstan	122	3,207	0.0%
Lithuania	68	1,519	0.0%
Qatar	66	1,343	98.6%
Saudi Arabia	51	1,872	31.8%
United Arab Emirates	92	1,990	98.9%
United States	78	1,657	92.9%

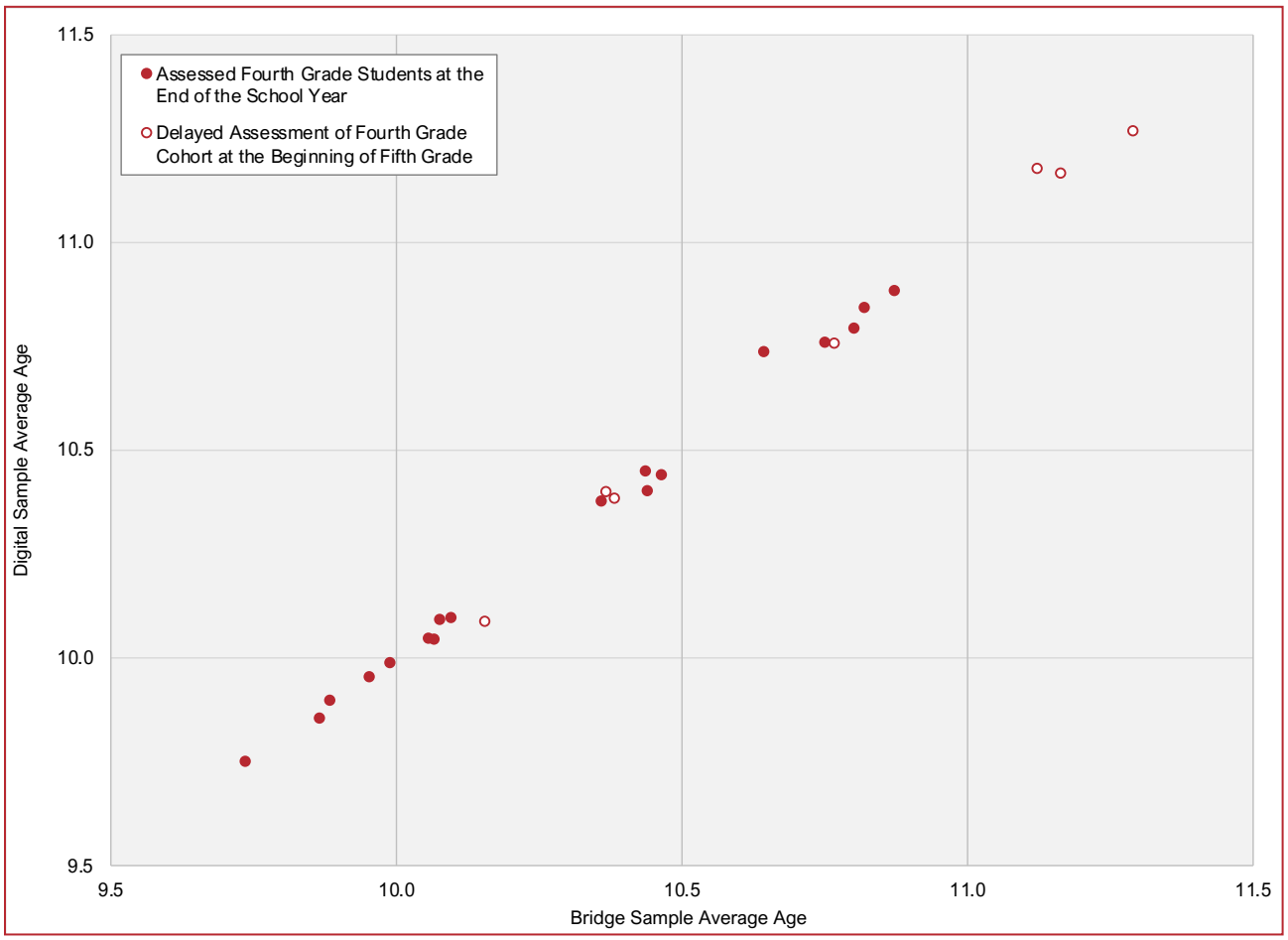
Comparison of Key Variables

Examining the equivalence of the bridge and digital samples also included the comparison of samples on key context variables. This section presents the results of several variables including students' age at the time of testing, the distribution of students by gender and language, and socioeconomic status. The results provided further validation of the equivalence of the bridge and digital samples. For all countries, outcomes on these variables can be considered equivalent for bridge and digital samples within the margin of error.

The students' average age for the bridge and digital samples are compared in Exhibit 12.13. The data points align well with the diagonal reference line for all the countries, whether assessed

in the fourth grade at the end of the school year or the fourth grade cohort at the beginning of fifth grade. There is very little dispersion from the diagonal reference line, which is evidence of the similarity of the average age of students between the two samples.

Exhibit 12.13: Plot of Averages Age of Students in Digital and Bridge Samples



In Exhibit 12.14, each of the bars show a country's difference between the percentage of girls in the digital sample and the percentage of girls in the bridge sample. Negative values indicate a larger percentage of girls in the bridge sample, and positive values indicate a larger percentage in the digital sample. The vertical lines represent 95 percent confidence intervals. With all confidence intervals crossing zero on the vertical axis, the difference in gender proportions between the two samples was non-significant in all countries. There was a slightly greater difference in the gender proportions in Malta. This is likely due to the relatively small bridge sample size in that country compared to the others.

Exhibit 12.14: Bar Graph of Differences in Percentages of Girls between Digital and Bridge Samples

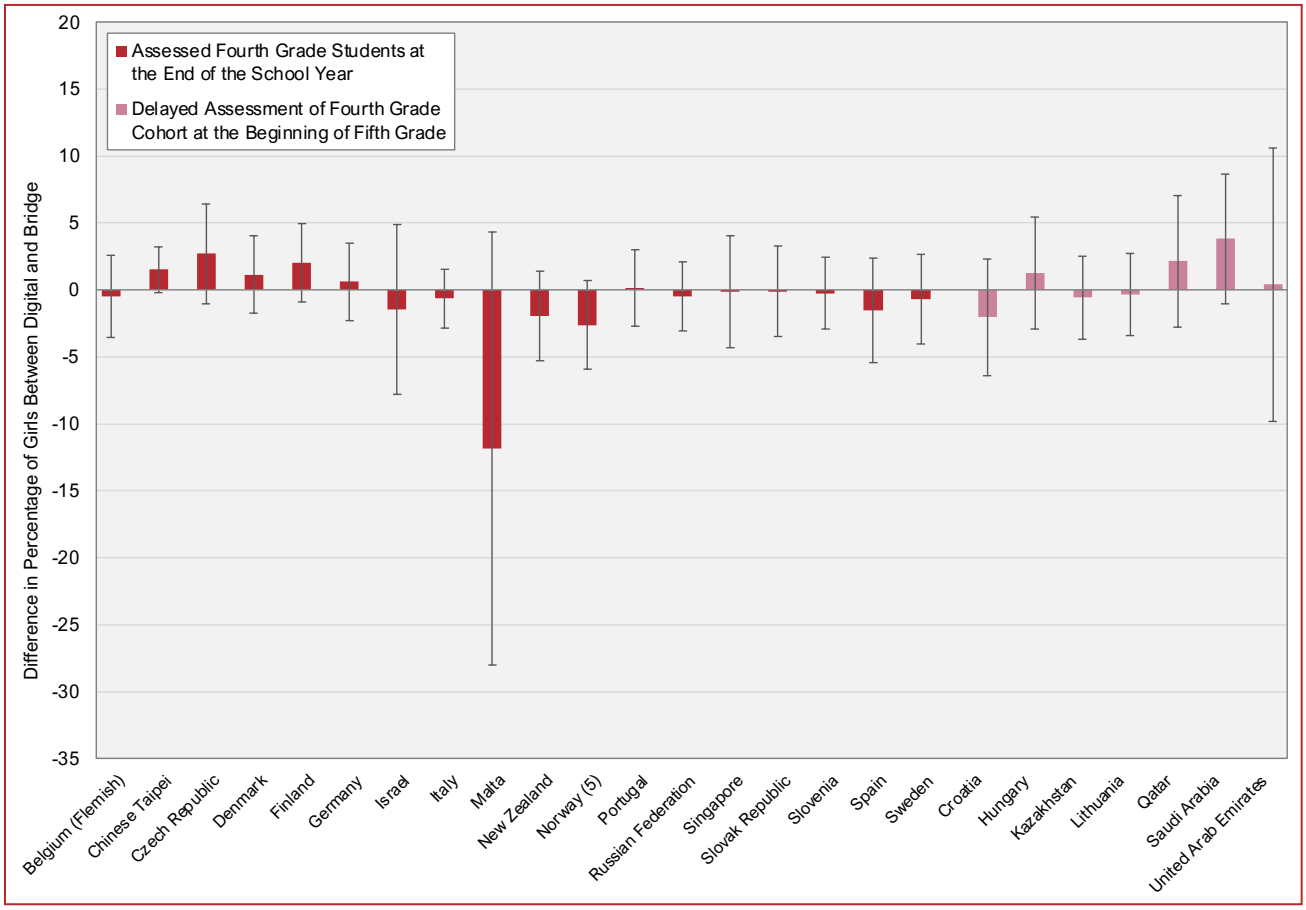


Exhibit 12.15 shows the percentage of students in each country by response category on the PIRLS 2021 Home Questionnaire item asking "how often does your child speak the language of the test at home?"—with response options "Always," "Almost always," "Sometimes," and "Never." There was a high level of agreement between the two sets of results, except for one data point (in red) belonging to Malta. As noted above, the smaller sample size in Malta can be understood as the reason for the larger uncertainty of the estimates.

Exhibit 12.15: Plot of Percentages of Students Speaking the Language of the Test at Home in Digital and Bridge Samples

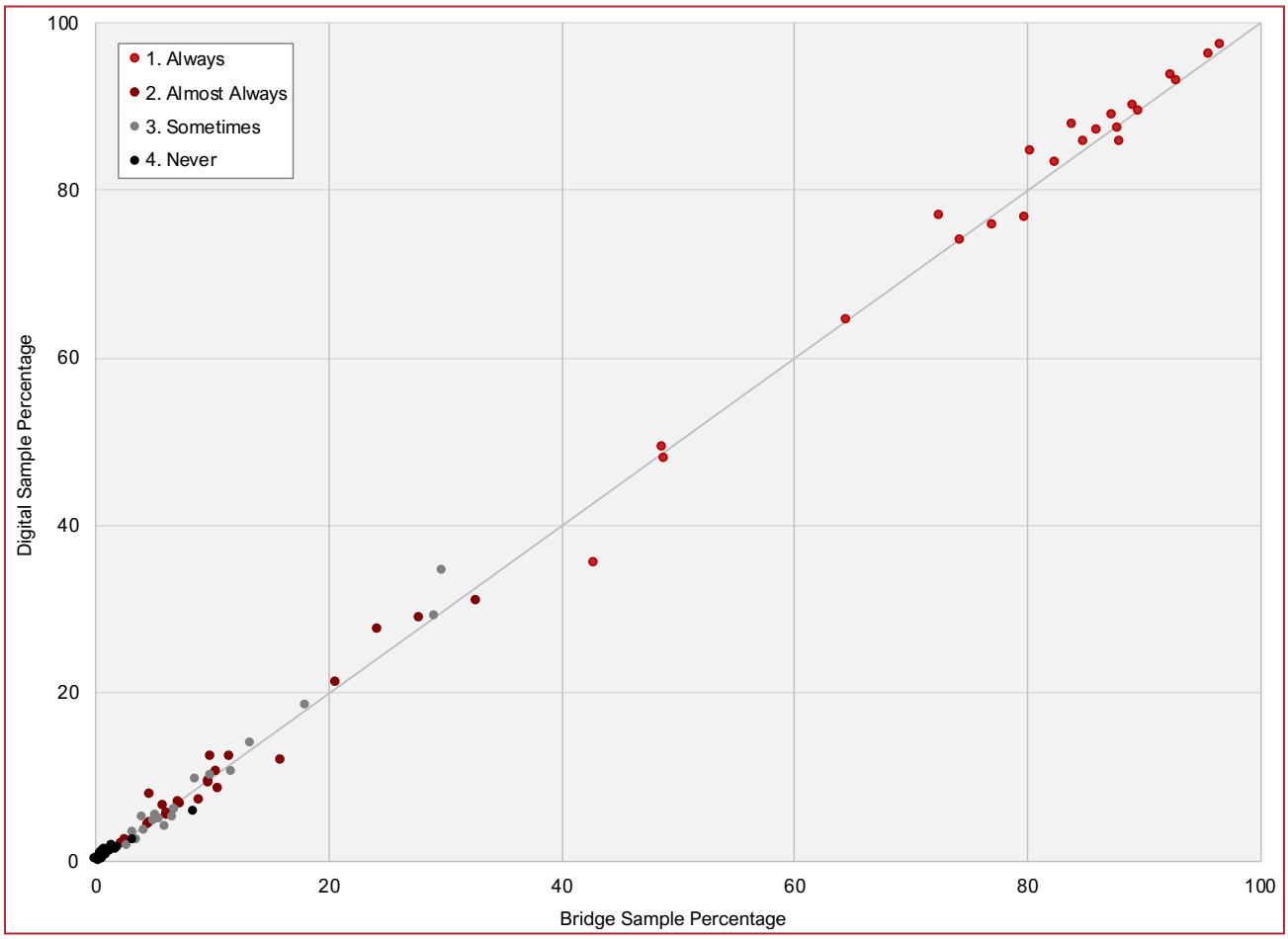
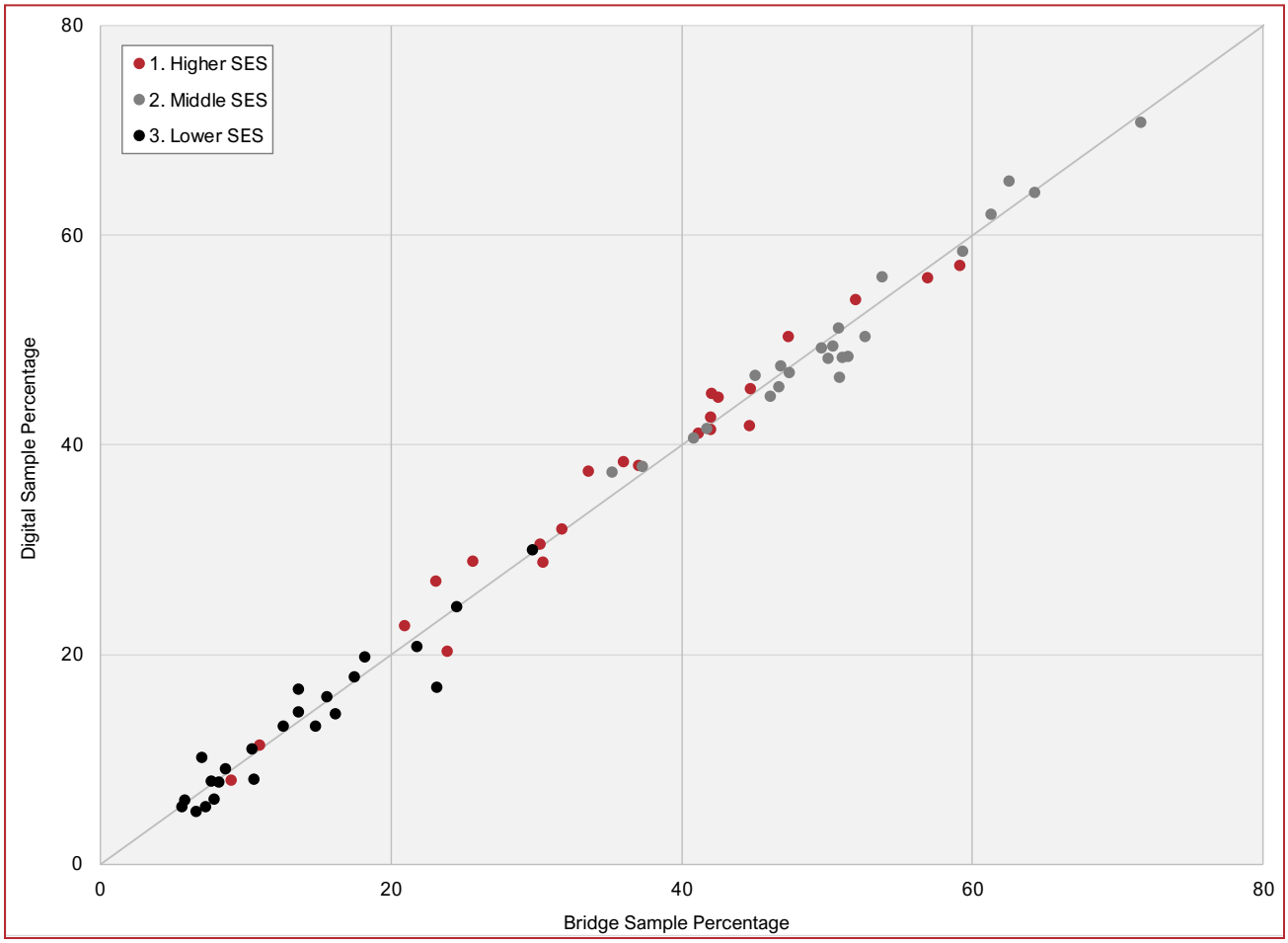


Exhibit 12.16 shows bridge-digital sample comparisons of the *PIRLS 2021 Home Socioeconomic Status* ("Home SES") context scale. Detailed information on this new context scale is provided in Chapter 15. Based on the scale scores, students were placed in one of three categories of socioeconomic status: higher, middle, or lower. For each country, the plot shows the percentage of students in each of the three categories for the bridge sample and the digital sample. The plot shows high consistency in the percentages between the bridge and digital samples across countries, except for one marginal outlier in the lower SES category for Croatia.

Exhibit 12.16: Plot of Percentages of Students by Home Socioeconomic Status in Digital and Bridge Samples



Summary of the Population-Based Linking Approach

As described in the subsections above, sizeable, heterogeneous differences were observed between average percent correct on trend items collected in paper-based bridge and digitalPIRLS samples. Consequently, PIRLS 2021 applied population-based linking using randomly equivalent samples to put the paper and digital assessment data on a common scale, capitalizing on the PIRLS 2021 data collection design without requiring assumptions of item equivalence (see full discussion in Chapter 10). The randomly equivalent samples drawn from the same populations established the link, accounting for the observed differences between the paper and digital assessments, and putting their item parameters on the same scale. Both the paper and digital assessments target the same reading construct and contain a large proportion of the same texts and items. Therefore, the population-based linking approach allows scale linking when both samples drawn from the same populations are equivalent in sampling quality and coverage of their respective populations.

Comparing sampling outcomes and key demographic variables showed that the data from the bridge and digital samples support the random equivalence assumption.

Linear transformations were used to put the PIRLS 2021 paper and digital data on the 2016 PIRLS trend reporting metric. The digitalPIRLS data were transformed by aligning the pooled ability distribution of the digitalPIRLS samples with the pooled ability distribution of the bridge samples, which already was transformed to the PIRLS trend metric using the concurrent calibration of the 2016 and 2021 paper data, as described in Chapter 11.

Country-Level Differences in PIRLS 2021 Average Scale Scores (Plausible Values) by Mode of Administration

With digital and bridge achievement results being on the same PIRLS reporting metric, the results can be directly compared. Observed country-level differences in achievement between the bridge and digital samples may result from sampling variation and country-specific effects after the international difference between bridge and digital sample achievement was accounted for through the equivalent groups linking.

This section compares the average scale scores between the paper bridge and computer-based assessments for all digitalPIRLS countries included in the PIRLS 2021 International Database. The first part compares the scale scores derived from the bridge samples to the computer-based scale scores for students administered one of the 18 digitalPIRLS booklets that have paperPIRLS equivalents. This first set of digital scores was derived solely based on the trend items and thus unaffected by the potential influence of the new digital and ePIRLS items. This comparison shows country-level residual differences in achievement when the same pools of paper-based and computer-based items are considered.

The second part compares the achievement scores from the bridge sample to the final computer-based achievement scores for all students in the digital sample, based on all digitalPIRLS and ePIRLS items. This comparison examines the final country-level residual differences in achievement when new digitalPIRLS and ePIRLS items were added to the assessment.

Average Scale Scores of Bridge and Digital Data Based on Trend Items Only

Exhibit 12.17 shows the average scale scores from the bridge and digital samples (derived based on trend items and used for validation purposes only) as well as the difference between them, together with their standard errors. After the international bridge-digital difference was accounted for through the population-based linking, there was almost no achievement difference between the averages of the bridge and digital samples in the first panel, based on trend items only. However, the average digital scale score across the countries with delayed testing and samples of older students is higher than the corresponding average bridge scale score, as shown in the second panel of the exhibit.

Exhibit 12.17: PIRLS 2021 Average Scale Scores for Paper Bridge and digitalPIRLS (Trend Items only) and Their Differences

Country	Bridge Average Score (Trend Items)	Digital Average Score (Trend Items)	Difference	
Assessed Fourth Grade Students at the End of the School Year				
Belgium (Flemish)	503 (3.6)	508 (2.7)	6 (4.5)	
Chinese Taipei	559 (3.2)	546 (2.1)	-14 (3.8)	▼
Czech Republic	545 (3.2)	539 (3.2)	-6 (4.6)	
Denmark	539 (3.6)	541 (2.6)	2 (4.4)	
Finland	544 (4.4)	548 (2.9)	4 (5.3)	
Germany	532 (3.6)	527 (2.8)	-5 (4.6)	
Israel	509 (4.1)	506 (2.6)	-3 (4.9)	
Italy	545 (3.4)	534 (2.3)	-10 (4.0)	▼
Malta	499 (7.9)	506 (3.3)	7 (8.5)	
New Zealand	517 (4.3)	519 (2.9)	3 (5.1)	
Norway (5)	523 (3.6)	531 (2.6)	8 (4.4)	
Portugal	531 (3.4)	520 (2.7)	-11 (4.4)	▼
Russian Federation	575 (4.6)	567 (3.9)	-7 (6.0)	
Singapore	582 (4.8)	584 (3.4)	2 (5.9)	
Slovak Republic	527 (5.7)	530 (3.1)	3 (6.5)	
Slovenia	520 (3.6)	526 (2.8)	6 (4.6)	
Spain	513 (3.9)	514 (2.7)	2 (4.7)	
Sweden	536 (4.3)	542 (2.9)	6 (5.2)	
Average (18)	**533 (1.0)**	**533 (0.7)**	**0 (1.2)**	
Delayed Assessment of Fourth Grade Cohort at the Beginning of Fifth Grade				
Croatia	553 (4.8)	556 (3.5)	3 (6.0)	
Hungary	542 (5.1)	542 (3.5)	0 (6.2)	
Kazakhstan	505 (3.9)	503 (3.1)	-2 (5.0)	
Lithuania	542 (4.2)	553 (2.7)	11 (5.0)	▲
Qatar	482 (7.8)	480 (3.8)	-2 (8.7)	
Saudi Arabia	417 (8.6)	452 (4.0)	34 (9.5)	▲
United Arab Emirates	468 (9.2)	479 (2.2)	11 (9.5)	
Average (7)	**501 (2.5)**	**509 (1.2)**	**8 (2.8)**	**▲**

( ) Standard errors appear in parentheses. Because of rounding some results may appear inconsistent.

▲ indicates digitalPIRLS performance significantly higher than bridge (α = 0.05)

▼ indicates digitalPIRLS performance significantly lower than bridge (α = 0.05)

Exhibit 12.17 shows a few statistically significant differences between paper and digital after accounting for the international mode difference. Differences varied in direction and magnitude across countries. However, countries that assessed fourth grade students at the end of the school year had, on average, the same performance on digitalPIRLS, while countries that assessed older students, the fourth grade cohort at the beginning of fifth grade, mostly had higher performance on digitalPIRLS. Countries who tested on the delayed schedule are displayed in the exhibits of the international report using a different color from countries that were able to test their students at the target age and grade.

The PIRLS country scale score differences are highly consistent with the country differences in average percent correct statistics (r = 0.90) in Exhibits 12.3 and 12.4, which provided an unscaled estimate of the country-specific bridge-digital differences. The slight deviations between the percent correct differences and scale score differences among countries are due to the non-linearity between the percent correct statistics and IRT-based scale scores, which can be understood when recalling the non-linearity of IRT item functions and considering that multiple non-linear item functions are involved in estimating a scale score from item response data (see Chapter 10). The effects of non-linearity are present already in the Rasch model and are even more pronounced with the application of two- and three-parameter IRT models, as are used in PIRLS.

Average Scale Scores of Bridge and Digital Data Based on All Items

To evaluate the extent of country-specific differences when all items, including ePIRLS items, are included, the final scale scores derived from the full digital samples were compared to the scores from the bridge sample for each country. The scale scores and the differences between the two samples are shown in Exhibit 12.18.

Exhibit 12.18: PIRLS 2021 Average Scale Scores for Paper Bridge and digitalPIRLS (All Items) and Their Differences

Country	Bridge Average Score (Trend Items)	Digital Average Score (All Items)	Difference	
Assessed Fourth Grade Students at the End of the School Year				
Belgium (Flemish)	503 (3.6)	511 (2.3)	8 (4.3)	
Chinese Taipei	559 (3.2)	544 (2.2)	-16 (3.9)	▼
Czech Republic	545 (3.2)	540 (2.3)	-6 (4.0)	
Denmark	539 (3.6)	539 (2.2)	0 (4.2)	
Finland	544 (4.4)	549 (2.4)	5 (5.0)	
Germany	532 (3.6)	524 (2.1)	-8 (4.2)	
Israel	509 (4.1)	510 (2.2)	1 (4.7)	
Italy	545 (3.4)	537 (2.2)	-8 (4.0)	
Malta	499 (7.9)	515 (2.7)	16 (8.3)	
New Zealand	517 (4.3)	521 (2.3)	5 (4.9)	
Norway (5)	523 (3.6)	539 (2.0)	16 (4.1)	▲
Portugal	531 (3.4)	520 (2.3)	-11 (4.1)	▼
Russian Federation	575 (4.6)	567 (3.6)	-7 (5.8)	
Singapore	582 (4.8)	587 (3.1)	5 (5.8)	
Slovak Republic	527 (5.7)	529 (2.7)	2 (6.3)	
Slovenia	520 (3.6)	520 (1.9)	-1 (4.0)	
Spain	513 (3.9)	521 (2.2)	9 (4.5)	
Sweden	536 (4.3)	544 (2.1)	8 (4.8)	
Average (18)	**533 (1.0)**	**534 (0.6)**	**1 (1.2)**	
Delayed Assessment of Fourth Grade Cohort at the Beginning of Fifth Grade				
Croatia	553 (4.8)	557 (2.5)	3 (5.4)	
Hungary	542 (5.1)	539 (3.4)	-2 (6.2)	
Kazakhstan	505 (3.9)	504 (2.7)	-2 (4.7)	
Lithuania	542 (4.2)	552 (2.3)	10 (4.8)	▲
Qatar	482 (7.8)	485 (3.7)	3 (8.6)	
Saudi Arabia	417 (8.6)	449 (3.6)	31 (9.3)	▲
United Arab Emirates	468 (9.2)	483 (1.8)	15 (9.4)	
Average (7)	**501 (2.5)**	**510 (1.1)**	**8 (2.7)**	**▲**

( ) Standard errors appear in parentheses. Because of rounding some results may appear inconsistent.

▲ indicates digitalPIRLS performance significantly higher than bridge (α = 0.05)

▼ indicates digitalPIRLS performance significantly lower than bridge (α = 0.05)

As shown in Exhibit 12.18, the scale score differences between the paper bridge and digital samples with all items included were generally consistent with those in Exhibit 12.17. On average, there was no achievement difference between the two samples across countries that assessed students at the end of fourth grade, as shown in the first panel. In contrast, the average digital scale score across countries with delayed test administration was higher than the corresponding average bridge scale score, as shown in the second panel. Although most scale score differences between the digital sample and the bridge sample did not change much when compared to the score differences in Exhibit 12.17, there were some changes, mostly small, with a few exceptions such as Norway (5), Malta, Spain, and Slovenia. These small changes reflect the effects of different numbers of items, the presence of new assessment blocks in the full data, and sampling variations.

Conclusion

PIRLS 2021 implemented the randomly equivalent samples design as PIRLS transitioned from paper-based to digital-based assessment. For each country that participated in digitalPIRLS, an additional sample of students, randomly sampled from the same population as the full digitalPIRLS sample, was administered a set of paper bridge booklets containing trend material. The average percent correct statistics across trend items for the paper bridge and digitalPIRLS data showed a sizeable international difference between the two testing modes, favoring paper-based performance. The PIRLS 2021 data collection was particularly complex. The school disruptions and delayed testing due to COVID-19 added to the complexity of the mode transition, which included changing the presentation of texts and items developed for PIRLS Literacy, and resulted in heterogeneous differences in item functioning between modes. Therefore, PIRLS 2021 adopted an equivalent samples-based linking approach to link the digital data to the PIRLS trend scale that required fewer assumptions. This approach does not assume there is a uniform mode effect and allows for each item administered in the digital assessment to have a set of parameters that differ from the paper-based assessment. Chapter 10 includes a more detailed description of the population-based linking approach and its advantages for use in PIRLS 2021.

The present chapter not only provided an overview and rationale for the population-based linking utilized in PIRLS 2021, but also showed how to describe and report differences between paper-based bridge and digitalPIRLS data with an understanding of how they relate to the PIRLS 2021 achievement results. This allows countries and secondary users of the data to evaluate the extent of observed differences in their PIRLS 2021 data at the country level while accounting for the limitations of the sample size of the bridge data. It is also intended to encourage researchers to examine how country samples may vary due to differences between paper bridge and digitalPIRLS responses, between on-time and delayed test administrations, between trend and new items, and among different subpopulations of students.

References

Bock, R. D., & Zimowski, M. F. (1997). Multiple group IRT. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433-448). New York, NY: Springer. https://doi.org/10.1007/978-1-4757-2691-6_25

Box, G. E. P., Hunter, J. S., & Hunter, W. G . (2005). *Statistics for experimenters: Design, innovation, and discovery, Second Edition*. Hoboken, NJ: John Wiley & Sons, Inc.

Clinton, V. (2019). Reading from paper compared to screens: A systematic review and meta-analysis. *Journal of Research in Reading, 42*(2), 288–325. https://onlinelibrary.wiley.com/doi/epdf/10.1111/1467-9817.12269

Dorans, N. J., & Puhan, G. (2017). Contributions to score linking theory and practice. In R. Bennett, & M. von Davier (Eds.), *Advancing human assessment. Methodology of educational measurement and assessment*. Springer, Cham. https://doi.org/10.1007/978-3-319-58689-2_4

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7, 1–26.

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), 1-23. https://doi.org/10.1186/s40536-018-0064-z

Foy, P., Fishbein, B., von Davier, M., & Yin, L. (2020). Implementing the TIMSS 2019 scaling methodology. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (pp. 12.1-12.146). Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods/chapter-12.html

Foy, P., Galia, J., & Li, I. (2008). Scaling the data from the TIMSS 2007 mathematics and science assessments. In J. F. Olson, M. O. Martin, & I. V. S. Mullis (Eds.), *TIMSS 2007 technical report*. Chestnut Hill: TIMSS & PIRLS International Study Center, Boston College. https://timssandpirls.bc.edu/timss2007/techreport.html

Gonzalez, E. J. (2003). Scaling the PIRLS reading assessment data. In M. O. Martin, I. V. S. Mullis, & A. M. Kennedy (Eds.), *PIRLS 2001 technical report*. Chestnut Hill, MA: Boston College. https://timssandpirls.bc.edu/pirls2001i/PIRLS2001_Pubs_TR.html

Haberman, S. J. (2015). Pseudo-equivalent groups and linking. *Journal of Educational and Behavioral Statistics, 40*(3), 254–273. https://doi.org/10.3102/1076998615574772

Jerrim, J., Micklewright, J., Heine, J.-H., Salzer, C., & McKeown, C. (2018). PISA 2015: how big is the 'mode effect' and what has been done about it? *Oxford Review of Education, 44*(4), 476-493. https://doi.org/10.1080/03054985.2018.1430025

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices* (3rd ed.). New York: Springer Science+Business Media.

Kong, Y., Seo, Y. S., & Zhai, L. (2018). Comparison of reading performance on screen and on paper: A meta-analysis. *Computers & Education, 123*, 138–149 https://doi.org/10.1016/j.compedu.2018.05.005

Millsap, R. E. (2011). *Statistical approaches to measurement invariance*. New York, NY: Routledge/Taylor & Francis Group.

von Davier, M. (2020). TIMSS 2019 scaling methodology: Item Response Theory, population models, and linking across modes. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (pp. 11.1-11.25). Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods/chapter-11.html

von Davier, M., Foy, P., Martin, M. O., & Mullis, I. V. S. (2020). Examining eTIMSS country differences between eTIMSS data and bridge Data: A look at country-level mode of administration effects. In M. O. Martin, M. von Davier, & I. V. S. Mullis (Eds.), *Methods and procedures: TIMSS 2019 technical report* (pp. 13.1-13.24). Boston College, TIMSS & PIRLS International Study Center. https://timssandpirls.bc.edu/timss2019/methods/chapter-13.html

von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & 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*(6), 671–705.

von Davier, M., & von Davier, A. A. (2007). A unified approach to IRT scale linking and scale transformations. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 3*(3), 115-124.

Appendix 12A

Comparing Country Level Average Percent Correct to the International Average

Consider the international average of a statistic, for example, an average percent correct across several items. In our case, these are the items that were designed for a paper-based assessment, PIRLS 2016, and still used in PIRLS 2021 as trend items for the paper-based assessment, and re-implemented for computer delivery for countries that chose to use the digitalPIRLS assessment.

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

$$\mu_I = \frac{1}{N}\sum_{k=1}^{N}\mu_k.$$

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 average percentages for country k are the weighted estimates of the percent 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 average percent 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

$$E\left(\widehat{M}_I\right) = E\left(\frac{1}{N}\sum_{k=1}^{N}\widehat{M}_k\right) = \mu_I$$

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

$$E\left(\widehat{M}_k\right) = \mu_k$$

and denote the associated standard errors by $\widehat{S}_k$. 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

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

with

$$cov\left(\widehat{M}_I, \widehat{M}_k\right) = cov\left(\frac{1}{N}\sum_{k=1}^{N}\widehat{M}_k, \widehat{M}_k\right) = \frac{1}{N}\,cov\left(\widehat{M}_k, \widehat{M}_k\right) = \frac{1}{N}\widehat{S}_k^2.$$

Plugging this result into the estimate provides

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

which is well defined whenever there are at least two countries, i.e., whenever $N \geq 2$. This is an application of estimating standard errors for comparisons against the international average, as described in Chapter 13.

Country Mode Differences, Corrected for International Mode Differences

The international estimate and the expected values of percent correct across paper items ("P" samples) will be denoted by

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

and the mean of percent correct across digital items ("D" samples) is

$$E\left(\widehat{M}_{ID}\right) = \mu_{ID}$$

Similarly, we have associated standard errors for the estimate of the international percent correct for paper, $\widehat{S}_{IP}$, and digital, $\widehat{S}_{ID}$, respectively, as we have for the country-level estimates $\widehat{S}_{kP}$ and $\widehat{S}_{kD}$. These can be calculated separately using Jackknife procedures (Chapter 13) and defined as given above. The bridge and the digital samples do provide an estimate of the mode difference

$$\Delta_{P-D} = \mu_{IP} - \mu_{ID} \tag{12.1}$$

at the international level. This mode difference is being controlled for in the linking design that uses the bridge and digital samples in a customary equivalent groups approach. That means the international difference is no longer relevant and can be taken out of country-level comparisons of achievement results between modes. Only remaining differences at the country level are relevant, as the overall difference is no longer affecting the plausible values that are provided in the international database. The international average of percent correct differences is already taken care of by the international adjustment. Consequently, the difference

$$\widehat{d}_{Pk-Dk} = \widehat{d}_{Pk} - \widehat{d}_{Dk} \approx \widehat{M}_{Pk} - \widehat{M}_{Dk} - \Delta_{P-D}$$

quantifies the relative paper versus digitalPIRLS difference in the average percent correct not accounted for by the international linking in the bridge study. For this estimated difference, we can calculate the standard error

$$\widehat{S}_{Pk-Dk}=\sqrt{(\widehat{S}_{Pd(k)})^2+(\widehat{S}_{Dd(k)})^2} \tag{12.2}$$

using the estimates defined as above

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

and

$$\widehat{S}_{Dd(k)}=\sqrt{\widehat{S}_{ID}^2+\left[\frac{N-2}{N}\right]\widehat{S}_{kD}^2}.$$

Note that these are almost the same as the standard error for the country mean average percent correct for paper versus digitalPIRLS, calculated separately. This statistic is adjusted by the standard error for the international percent correct (separately calculated by mode) but adjusted for the number of countries included in the international mean.

Achievement Data Comparisons based on Bridge and Digital Samples

The comparison, once the linking is accomplished, is rather straightforward. The standard error estimates for the bridge sample averages and the digitalPIRLS averages can be used to calculate the standard error of the difference for countries where schools were selected to test either using the paper bridge or the digitalPIRLS 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.

Assuming independent samples from the same population, the mean difference between paper bridge sample (P) and digital sample (D) in country k, as given in equation (12.1), can be evaluated using the standard error of the difference for independent samples, given in equation (12.2). However, this is no longer appropriate and may overestimate the standard error if students were assigned to paper or digitalPIRLS 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 digital 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. A third set of countries would have some schools that assign one class to bridge and another to digitalPIRLS, and the 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.

For the exhibits in this chapter, for simplicity of exposition, we assume independent samples of students taking the paper bridge and the digital assessment.