Annex H
New Procedures for PISA 2018 Population Modelling

This Annex describes four procedural changes introduced to the PISA 2018 population modelling. These procedural changes were implemented to 1) address situations in which one or more subpopulations in a country/economy are oversampled, 2) incorporate process data, specifically response time (RT) variables, in the population modelling[1], 3) model the financial literacy sample data, and 4) improve the modelling of the Une Heure (UH) cases.

Before operationalising these changes, experimental and simulation studies were conducted for each change to assure that the new procedures would improve measurement accuracy without harming comparability or the measurement of trends. In the sections below, we explain the rationale behind these studies, describe their design, and present their results. It is important to note that these four changes were implemented in generating the plausible values (PVs) reported in the public file for the PISA 2018 main survey data. Here is a brief summary of each of the changes.

1. *Use of School-Level Information as Direct Covariates*. Because schools are the primary sampling units, it is important to reflect between-school variations in the population modelling. In PISA 2015, this was achieved by including contrast codes of school identifiers (school ID) in the principal components analysis (PCA) along with other background questionnaire (BQ) variables. However, when a country/economy oversamples a subpopulation of interest for special reporting, or the oversampled subpopulation is much smaller than the country/economy's overall population—resulting in a very small contribution to the overall population estimate after being weighted—the school information for the subpopulation may not be taken into account as much as it should. After studying a number of alternative approaches using data from Northern Ireland (as part of the United Kingdom [UK] data) as an example, the preferred approach was to use the leave-one-out domain-specific school-level weighted likelihood estimates (WLEs) as continuous covariates in the multivariate latent regression model. This approach, which replaced all the contrast-coded school ID variables with school-level WLEs for each domain, worked well in all cases with oversampling. In 2018, it was used for the modelling of the main sample data (11 countries/economies) and the financial literacy sample data (four countries/economies) where a significant variation in sampling rates existed.

[1]] Note that two item response time variables are reported in the public use files (PUF): one reported using the item ID with the suffix T (T variable), and one using the item ID with the suffix TT (TT variable). Considering a student could come back to revisit an item after visiting other items in the unit, the T variable captures the time spent during the last visit to the item alone, whereas the TT variable captures the aggregate time across all visits to the item. This last variable provides a better accounting of the total time a student may have spent responding to the item. However, because the TT variable was not available for the final scaling of the 2018 Main Survey data, results presented in this Annex as well as the final 2018 Main Survey plausible values and the group proficiencies statistics reported are based on the T variable. Recent population modeling re-analyses using the TT variable have shown that plausible values, group proficiency estimates obtained using T or TT are equivalent as differences are well within imputation error.

2. *Use of Response Time Information as Conditioning Variables.* Since the implementation of computer-based assessment in 2015, cognitive item-level RT data have been newly available to the public. This information is getting more interest from researchers and practitioners who conduct secondary analyses using the RT data and PVs. To support making inferences about the relationship between RT and PV, it is important to incorporate the RT information in the process of generating PVs without introducing bias. Furthermore, incorporating the RT information in the population modelling contributes to the increase in measurement precision of proficiency. The approach implemented in 2018 utilised test-level variables aggregated from the item-level RT variables, taking into account item type and assessment hour (first or second hour). These new variables were contrast coded and included in the PCA along with other BQ variables for all CBA countries/economies where RT information was available.
3. *Modelling the Financial Literacy Sample.* Unlike in PISA 2015, in 2018, the financial literacy domain was administered to, and analysed as, a separate sample from the main sample. In each of the 21 countries/economies that collected financial literacy samples, an additional sample of students was selected and administered the financial literacy domain and either the mathematics or reading domain. For more effective analyses and modelling of the financial literacy sample data, students from the main sample who took forms assessing both reading and mathematics were combined and analysed together with the financial literacy sample.
4. *Modelling Large Proportions of UH Cases.* The number of UH cases has grown in some countries/economies, requiring some improvement in their incorporation in the population modelling and generation of PVs. Given that UH cases are different in their representativeness of the target population and also different in the instruments they are administered (for both the cognitive and BQ variables), the feasibility of a new approach was studied. Instead of including the UH cases in a single conditioning model estimated in each country/economy as was done in 2015, a mixed approach was used. In this mixed approach, to generate PVs for the non-UH cases, the population model parameters were estimated using only the non-UH cases and the entire set of BQ questions (full model[2]). In contrast, to generate PVs for the UH cases, the population model parameters were estimated using the entire sample (including both the UH cases and non-UH cases), but only the subset of BQ questions administered to the UH students was included (reduced model). The full model is used to generate PVs for the non-UH students, and the reduced model is used to generate PVs for UH students – PVs for non-UH cases generated from the reduced model were discarded. For the reduced model, the UH indicator variable is specified as a direct dummy-coded covariate instead of being processed through the PCA. The latter was the method used in 2015. For the full model, the UH indicator variable is not used because the full model only involves non-UH cases. This new approach was implemented in 2018 for five countries/economies with a relatively large number of UH cases (more than 200 cases included in the main sample), and the 2015 approach was used for the other countries with a smaller number of UH cases.

As a reminder, the population modelling in PISA refers to the combination of item response theory and multivariate latent regression modelling (IRT-LRMs)—the latent abilities being regressed onto the BQ information. Operationally, students' latent abilities are estimated based on the

[2] Note that 'full model' in the 2018 approach is different from the 2015 approach. The 2015 approach used the complete dataset including UH cases and non-UH cases for estimating latent regression parameters, while the 'full model' in 2018 approach used only the non-UH cases for estimating latent regression parameters. Both the 'full model' in the 2018 approach and the 2015 approach use the entire set of BQ questions.

cognitive item responses, and the item parameters are fixed to the estimates obtained from the unidimensional IRT scaling stage for each domain (von Davier & Sinharay, 2014). To accommodate the large numbers of BQ variables available from the international and country/economy-specific BQ, PCA is conducted as a variable reduction technique, and the extracted principal components (PCs) are used as conditioning variables in the model. This approach has a long history in large-scale assessment and is often referred to as principal components regression (Jolliffe, 1982).

Since PISA 2015, the IRT scaling has used all international data, and both international and country-by-language group-specific item parameters have been estimated. Then, a multivariate latent regression model is fitted for each country/economy. The number of PCs to be included in the latent regression model is determined by the number of PCs that explain 80% of the total BQ variance for the country/economy, or 5% of the raw student sample size per country/economy, whichever is less. These criteria ensure that the estimated model captures a large amount of information in the conditioning variables, but at the same time, avoids overparameterization of the model, which could lead to unstable outcomes and erroneous inferences (OECD, 2017). With these criteria in place, a large but limited number of extracted PCs can be specified as covariates in the multivariate latent regressions.

For PISA 2018, it is important to note that the multistage adaptive testing (MSAT) routing path information is not used as an additional variable to define the groups used in the multiple-group IRT models or as a covariate for student characteristics in the multivariate latent regressions. This decision is based on theoretical research (Glas, 1988; Jewsbury, Lu, & van Rijn, 2019) and simulation studies (van Rijn & Shin, 2019).[3] Because routing decisions in PISA are largely based on cognitive responses (i.e., sum scores based on the machine-scored items), using this information again as covariates in the multivariate latent regression would violate the conditional independence assumptions underlying the IRT-LRMs. More details about the PISA 2018 reading MSAT designs and related outcomes and considerations can be found in Chapters 2, 9, and 12 of the technical report and in an OECD working paper (Yamamoto, Shin, & Khorramdel, 2019).

Change 1: Use of School-Level Information as Direct Covariates

As stated above, PCA is used as a variable-reduction technique as a preliminary step in the population modelling. Using PCA allows for the accommodation of the large number of contextual variables collected in the PISA BQ. Previous studies have found that PCA performs well as a dimension-reduction technique in group-level reporting situations (e.g., Mislevy, 1991).

However, Jolliffe (1982) explicitly stated that principal components regression analysis can have a problem when only the PCs with the highest eigenvalues are retained. That is, important information may still exist in the PCs with smaller eigenvalues. This possibility was argued by Benton (2019) who showed that PCA could potentially lead to bias in parameter estimation for

[3] Note that the Programme for the International Assessment of Adult Competencies (PIAAC) takes different approaches than PISA (OECD, 2013), although PIAAC implements the MSAT designs for all domains as well. The major difference is that PIAAC uses the MSAT routing path information as covariates in the multivariate latent regression because external variables (e.g., education level and native-versus-nonnative speaker) in addition to the cognitive item responses are used in the routing decision.

some subgroups of students with relatively small sampling weights, as the PCA is conducted using weighted data. If a subgroup or subgroups were not prevalent in the full population, information specific to the subgroups (e.g., regional differences) would not be retained.

For example, in PISA 2015, students from Northern Ireland (NIR) were deliberately oversampled as part of the UK main sample, and 2,401 students from NIR participated in PISA 2015 (17% of the UK's sample). When the national sampling weights in the PISA data were used, NIR's students were estimated to constitute only 3% of the UK's 15-year-old school population.[4] Therefore, after weighting, the variable indicating whether a student attended a school in NIR could only account for an extremely small proportion of the total variance, and therefore, the PCA assigned little priority to retaining this information. To address this issue, an alternative approach was proposed and studied to retain the uniqueness of the schools. This alternative approach was to use the aggregated school-level information as direct continuous covariates in the multivariate latent regression models (instead of first processing contrast codes of school IDs through the PCA). Note that in this alternative approach, all the remaining BQ variables were processed through the PCA, as was done in PISA 2015. The only difference in 2018 was that the aggregated school-level information was included as direct continuous covariates in the multivariate latent regression models, in addition to the PCs extracted from the PCA.

In the study designed to investigate this issue, the NIR data collected in the 2015 cycle were used to examine the feasibility of alternative approaches. As noted, the NIR data (N=2,401) were an oversample within the UK sample (N=11,046)[5]—NIR data represented about 17% of the total UK sample when unweighted, but only about 3% of the total UK sample when weighted. First, we investigated if a separate population model for NIR (i.e., treating NIR as a separate country/economy) could be a feasible option. To address this possibility, two alternative population modelling runs using only the NIR data were carried out: 1) with contrast-coded school IDs (same procedure as in PISA 2015), and 2) without school IDs. These two alternative runs can be considered to be "smaller" models due to the smaller sample size, which resulted in a fewer number of PCs based on the 5% sample size rule for the NIR-only data. In particular, the comparison between the first model (NIR data with school IDs) and the second model (NIR data without school IDs) was expected to indicate the impact of the uniqueness of the NIR schools. Moreover, the comparison of the first model (NIR data with school IDs) and the reported 2015 results were expected to show the impact of the uniqueness of the NIR schools in addition to the effect of having a smaller population model.

Table H.1 presents the results from the PCA for the reported 2015 results and the two alternative runs of smaller population models. Note that the number of PCs for the two alternative runs was based on the one-twentieth sample size rule, while the number of PCs for the reported 2015 UK results was based on the 80% explained variance rule.

Table H.1. Results from principal components analyses

	N	Number of PCs	% of Explained Variance

[4] Note that in PISA 2015, contrast codes of school IDs were included in the PCA, along with other BQ variables.

[5] Note that the UK sample excluded Scotland, which participates through a separate national centre, but included the NIR oversample.

			in BQ
Reported 2015 UK	11,046	505	80.01
(1) NIR with school ID	2,401	120	57.99
(2) NIR without school ID	2,401	120	62.86

As can be seen from Table H.1, the two population models that used only NIR data resulted in a much smaller proportion of explained variance in the BQ. Therefore, it is expected that the smaller population models would yield increased measurement errors compared to the larger population model that was used to report UK results in 2015. This is confirmed in terms of the standard deviation (SD) of the PVs reported in Table H.2. The table presents the summary of the PVs with respect to their means and SDs for each domain based only on the NIR data (N=2,401). Although the PV means seem quite similar across the three different population modelling strategies, except for the Collaborative Problem Solving (CPS) domain,[6] a comparison of the reported 2015 results to both of the smaller population models indicates that the SD of the PVs is much larger in the smaller population models—as much as 12–27% in the core domains (mathematics, reading, and science) and as much as 21% in the CPS domain.

Table H.2. Comparison of plausible value-based statistics under alternative modelling approaches for Northern Ireland only (2015)

NIR data only	Reported 2015		With School ID		Without School ID	
N=2,401	Mean (SE)	SD (SE)	Mean (SE)	SD (SE)	Mean (SE)	SD (SE)
Math	492.8 (4.6)	77.5 (2.0)	493.7 (4.7)	98.4 (4.7)	497.8 (4.8)	96.1 (4.4)
Reading	497.0 (4.6)	83.8 (2.0)	504.0 (5.7)	101.7 (3.0)	503.9 (5.5)	102.5 (4.6)
Science	500.1 (2.8)	89.8 (2.0)	500.6 (3.7)	100.4 (2.5)	499.1 (4.1)	99.9 (2.7)
CPS	514.0 (3.7)	88.1 (1.9)	521.5 (5.6)	106.5 (3.8)	518.2 (4.6)	106.2 (5.3)

Note: CPS = Collaborative Problem Solving

A comparison of the two smaller models (with school ID vs. without school ID) indicates the unique effect of having contrast codes of the school IDs, given the smaller population model that was used. Note that in the PCA, all contrast codes of the school IDs were included in the first modelling approach (NIR data with School IDs); thus, it was expected that the differences across schools would be reflected better in this approach than in the second modelling approach (NIR data without school IDs). When the results were compared, however, the effect of having school IDs seemed negligible in terms of the difference in PV means and SDs. Rather, the more important finding here is that smaller population models that used only NIR data resulted in much larger SDs.

From the comparison of the three models in Tables H.1 and H.2, for the overall reported results for NIR in 2015, the single larger population modelling combined with UK appears to be more precisely estimated with smaller measurement errors. As a next step, given that a larger population

[6] When the largest difference in the PV means among the three models was less than one standard error, the difference was considered to be insignificant and negligible. For example, in reading, (504-497)/sqrt(4.6^2+5.7^2) resulted in 0.96, which is less than 1. Thus, the PV means in reading across the three models were considered to be more precisely estimated.

model is more desirable, several alternative approaches were attempted and discussed at the Technical Advisory Group (TAG) meeting. Eventually, the final approach chosen was to use the leave-one-out (LOO) domain-specific school-level WLEs as direct covariates in the multivariate latent regression model (*School-Level WLEs*). More specifically, domain-specific school-level WLEs were assigned to individual students by averaging students' WLEs for the given school excluding that corresponding student (i.e., LOO). Unlike most of the BQ variables, these covariates were not processed as part of the PCA, but included in the multivariate latent regression model as direct continuous covariates along with the extracted PCs. Note that the use of school-level averages of the WLEs including the corresponding student would violate the conditional independence assumption (i.e., independence between the covariates and the item responses, conditional on the latent ability) required in IRT-LRMs (Jewsbury et al., 2019; Thomas, 2002). However, by excluding the corresponding student (i.e., LOO), this violation does not occur. In practice, LOO did not introduce any discernible differences in the school means in most cases.

Table H.3 presents comparisons between the two approaches using the same 2015 data from the UK (excluding Scotland, which participates through a separate national centre, but including the NIR oversample; N=11,046): 1) using contrast codes of the school IDs (2015 approach) and 2) using the continuous school-level WLEs (2018 approach). Note again that these results were from a single larger population model and were summarised for NIR and for the whole population of the UK. For the UK (including the NIR oversample; N=11,046), the PV-based means and SDs look quite similar between the two approaches. In addition, for the UK, the distances between the 5th and 95th percentiles look similar between the two approaches, the 2015 approach and 2018 approach. For the NIR oversample (N=2,401), the PV-based means looked similar between the two approaches as are seen from the UK sample. However, the PV-based SDs, as well as the distances between the 5th and 95th percentiles, were greater for the school-level WLEs approach. Using this UK and NIR example, it was inferred that the school-level WLEs allowed for a greater degree of between-school variation to be retained in the imputations for students in NIR.

Table H.3. Comparison of plausible values-based statistics under alternative modelling approaches (United Kingdom, 2015)

UK[7] (N=11,046)	Contrast codes of school IDs (2015 approach)				School-level WLEs (2018 approach)			
	Mean (SE)	SD (SE)	5th Percent.	95th Percent.	Mean (SE)	SD (SE)	5th Percent.	95th Percent.
Math	492.6 (2.7)	93.3 (1.4)	336.0 (4.5)	642.0 (4.1)	492.8 (2.6)	91.4 (1.3)	338.2 (3.9)	639.3 (4.0)
Reading	498.4 (3.0)	97.2 (1.2)	336.0 (4.7)	654.8 (4.5)	498.9 (2.7)	96.6 (1.3)	336.6 (5.2)	654.4 (4.1)
Science	510.3 (2.8)	100.0 (1.1)	345.2 (3.2)	671.8 (3.7)	510.2 (2.8)	99.6 (1.1)	345.1 (3.6)	670.5 (3.6)
CPS	519.7 (2.9)	103.3 (1.2)	348.1 (4.5)	687.2 (4.2)	519.9 (2.6)	102.4 (1.3)	348.2 (4.6)	684.6 (4.3)
NIR (N=2,401)	Contrast codes of School IDs (2015 approach)				School-level WLEs (2018 approach)			
	Mean	SD	5th	95th	Mean	SD	5th	95th

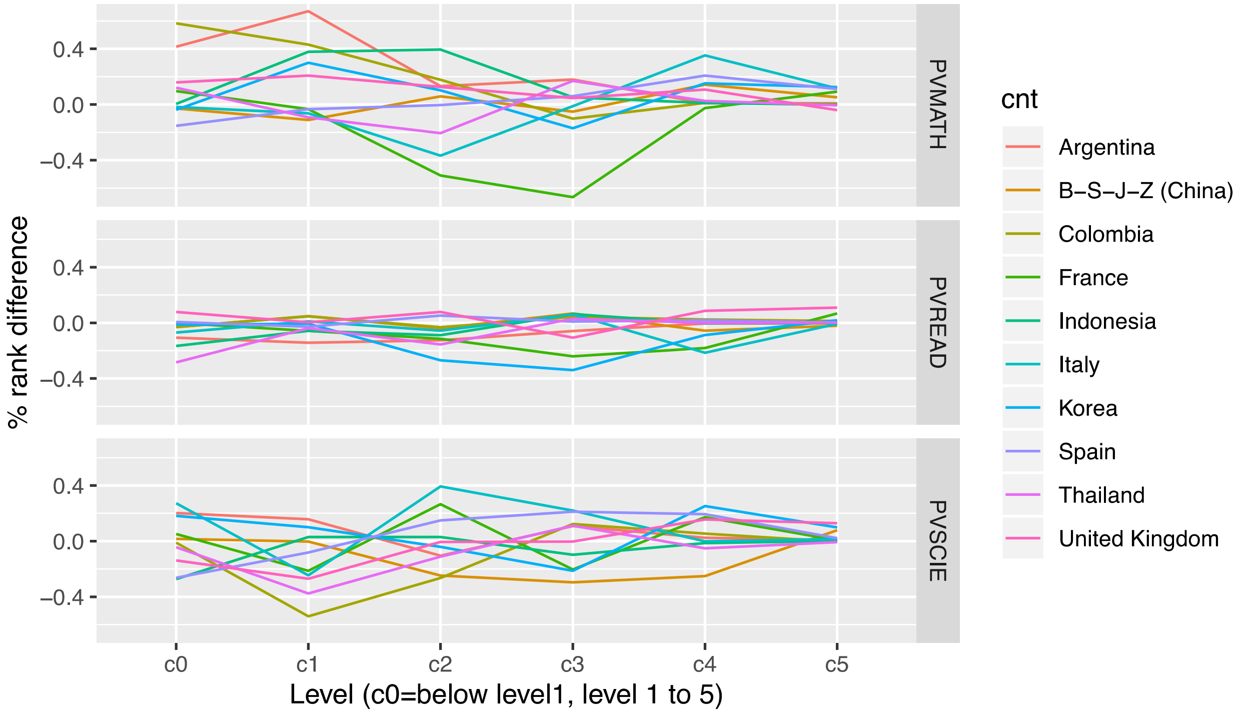
[7] Note that United Kingdom (UK) sample excluded Scotland, which participates through a separate national centre, but included the NIR oversample.

	(SE)	(SE)	Percent.	Percent.	(SE)	(SE)	Percent.	Percent.
Math	492.8	77.5	363.5	616.6	495.9	81.3	359.5	624.2
	(4.6)	(2.0)	(6.1)	(6.8)	(3.6)	(1.9)	(5.5)	(5.4)
Reading	497.0	83.8	355.8	632.0	497.1	86.4	351.6	636.0
	(4.6)	(2.0)	(7.0)	(6.8)	(3.5)	(2.1)	(5.7)	(5.5)
Science	500.1	89.8	352.3	644.3	500.5	91.6	349.6	648.7
	(2.8)	(2.0)	(4.8)	(4.6)	(3.1)	(1.9)	(4.8)	(4.2)
CPS	514.0	88.1	365.5	653.9	515.8	89.4	363.6	656.0
	(3.7)	(1.9)	(6.6)	(5.2)	(3.5)	(1.8)	(5.8)	(6.3)

Although the school-level WLEs approach appears promising for retaining between-school variation in the imputation with the example of NIR, it was practically impossible to accommodate this change for all participating countries/economies in the PISA 2018 main survey within the existing timeline. Thus, this new approach was only applied to the main sample from 11 countries/economies that significantly oversampled certain subpopulations for which they desired more detailed reports. Based on the ratio between the minimum and maximum of the student weights (i.e., minimum/maximum of the student weights < 0.01), the following countries/economies were selected to have the new 2018 approach applied: Argentina, B-S-J-Z (China), Colombia, France, Indonesia, Italy, Korea, the Russian Federation (and the three regions), Spain, Thailand, and the UK (excluding Scotland). For the six countries/economies that assessed global competence, a four-dimensional latent regression model (i.e., for mathematics, reading, science, and global competence) was applied, while for the five countries/economics that did not assess global competence, a three-dimensional latent regression model (i.e., for mathematics, reading, and science) was applied. This new approach was also applied when producing the reading subscale PVs, except for the one paper-based country/economy (Argentina), for which the reading subscales were not reported. Finally, this new approach was also applied to the financial literacy sample from four countries/economies (Indonesia, Italy, the Russian Federation, and Spain). For the rest of the countries/economies, the 2015 procedure was retained (i.e., contrast codes of the school IDs used in the PCA with other BQ variables, and the extracted PCs used as covariates in the multivariate latent regressions).

For the main sample from the 11 countries/economies for which the new school-level WLEs approach was applied, further comparisons were made to evaluate the differences with respect to the percentages of proficiency level changes between the 2015 approach and the new 2018 approach. Figure H.1 illustrates the differences in the percentages of proficiency levels (2018 approach – 2015 approach). It illustrates that the largest absolute difference is smaller than 0.8 percentage points across all domains and countries/economies, and for the major domain of reading, the range of differences is between –0.4 to 0.2 percentage points. Therefore, changing to the school-level WLEs approach for countries/economies that significantly oversampled certain subpopulations would only have marginal effects on the country/economy rankings and students' level proficiencies. However, in principle, and as illustrated with the NIR example, this new approach allows for a better description of the subgroup skill distributions (e.g., regional differences) by better reflecting between-school variations.

Figure H.1. Comparison of percentage of proficiency levels between 2015 approach and 2018 approach



Change 2: Use of Response Time Information as Conditioning Variables

The second change is a first step toward incorporating process data, in particular, RT information (i.e., time spent on the item), into the population modelling. RT information has much potential to contribute to data quality investigations, thereby enhancing the validity and reliability of the test results. For example, educational researchers have shown a keen interest in using RT and process data to evaluate the validity of cognitive responses and to provide insight into test-taking strategies, motivation, and engagement of both individuals and groups (e.g., Goldhammer, Martens, Christoph, & Lüdtke, 2016; Lee & Haberman, 2016; Lee & Jia, 2014; Meyer, 2010; van der Linden & Guo, 2008; van der Linden & Sotaridona, 2006; Weeks, von Davier, & Yamamoto, 2016).

For PISA, incorporating RT information into the IRT-LRMs is important for the following reasons. First, if the RT variables have not been used to generate the PVs but are available in the public database, secondary analyses estimating the relationship between the PVs and the RT variables may be biased (Meng, 1993; Mislevy, 1991). In addition, RT information may uniquely contribute to the prediction of proficiency and, therefore, improve the quality of the PVs and increase measurement precision (Mislevy, 1991; von Davier, Khorramdel, He, Shin, & Chen, 2019; Shin, Jewsbury, & van Rijn, 2019). In particular, two recent studies (Shin, Jewsbury, & van Rijn, 2019; Shin, Yamamoto, Khorramdel, Robin, von Davier, Gamble, & Zhao, 2019) revealed that incorporating RT information into the PISA population modelling is promising because the inclusion of RT variables resulted in a substantial increase in the measurement precision of proficiency (i.e., greater explained variance) with negligible differences in the regression coefficient estimates specified in the LRMs. Using PISA 2015 data, it was found that the gain in measurement precision was about 16%.

One remaining concern is the violation of the conditional independence assumption in the IRT-LRM because RT data is recorded at the item level; incorporating both the RT data and item responses for the same item, if correlated, can violate the conditional independence assumption of IRT. Although limited, the empirical study conducted by Shin, Jewsbury, and van Rijn (2019) revealed that the differences in the estimates of regression coefficients when RT was included or excluded are negligible, and the benefits of incorporating RT outweigh the potential bias from violating the conditional independence assumption. Therefore, experimental studies were conducted to evaluate the incorporation of RT information in the PISA 2018 population modelling through careful data processing of the RT variables. Another special consideration when processing the RT data is that item type should be considered in PISA with multiple languages in a mixed-format test (i.e., a mixture of multiple-choice items and constructed-response items; Yamamoto, 2019; Shin, Kerzabi, Joo, Robin, & Yamamoto, 2020). When the latent correlations between the RT scales of constructed-response and multiple-choice items were estimated, the mean and median of the correlations were both approximately .50, and the highest correlation was approximately .70. This implies that the RT scales measured by different item types are distinct, and each RT scale provides somewhat unique information that is not captured by the RT scale of another item format.

Therefore, following previous studies, the RT variables were treated as person covariates (e.g., proxy of working speed) in the population modelling of the PISA 2018 data. Item-by-person interactions were reduced by standardising and aggregating item-level RT variables within a country/economy. That is, for each item, the original continuous RT variable (recorded in milliseconds) was converted to deciles: Students who answered the item were ranked for each item, and these students were categorised into one of the ten ordered groups from the fastest to the slowest responders in each country/economy, creating a decile RT variable for each item. The item-level decile RT variables were then aggregated across items per assessment hour, regardless of the domain, based on two different approaches:

1. Summarise item-level RT deciles by hour (RT by hour)[8] so that each student receives four summarised RT variables. These four RT variables were 1) mean of the item-level RT deciles across items administered in Hour 1; 2) SD of the item-level RT deciles across items administered in Hour 1; 3) mean of the item-level RT deciles across items administered in Hour 2; and 4) SD of the item-level RT deciles across items administered in Hour 2.
2. Summarise the item-level RT deciles by item type and hour (RT by item type and by hour) so that each student receives eight summarised RT variables. Similar to the method above, RT data were aggregated again, this time including an additional factor of item type, as recommended by Shin et al. (2020). That is, item-level RT deciles were aggregated not only by hour, but also by item type (machine- vs. human-coded items[9]). The means and SDs of the item-level RT deciles were computed by hour and by item type, resulting in eight variables for individual students. Thus, each student has four

[8] The assessment was divided into two "hours" or sessions. During each of the sessions, the items for the major domain were administered (using the MSAT design) or two clusters from the minor domains were administered.

[9] For the analyses presented in this document and for the PISA 2018 population modelling, item types included machine- and human-coded items. All human-coded items were constructed response items, while only a few machine-coded items were constructed response items.

> RT variables for Hour 1: 1) mean of the item-level RT deciles across machine-coded items administered in Hour 1; 2) SD of the item-level RT deciles across human-coded items administered in Hour 1; 3) mean of the item-level RT deciles across machine-coded items administered in Hour 1; and 4) SD of the item-level RT deciles across machine-coded items administered in Hour 1. Another set of these four variables was created for the second hour, resulting in eight variables in total.

Means and SDs of the item-level RT deciles for each assessment hour (RT-by-hour approach) and for each assessment hour and by item type (RT-by-item-type-and-hour approach) were then further converted to deciles so that the categorical decile values could be contrast coded and included in the PCA, as with the other BQ variables. By taking the mean RT over items and then again over deciles, the relationship between RT and response at the item-level is almost nonexistent, which makes the chance of violating the conditional independence assumption remote.

For this experimental study, the following 11 countries/economies (covering a wide range of performance levels and cultures) were used: Australia (AUS), Brazil (BRA), Costa Rica (CRI), France (FRA), Germany (DEU), Greece (GRC), Japan (JPN), Korea (KOR), Mexico (MEX), the Netherlands (NLD), and the United States (USA). The two approaches above were implemented independently for the PCA, and subsequently, LRMs were fitted for each condition and compared to a baseline model that did not include any RT information (2015 approach).

Table H.4 presents the number of PCs retained for each model and the corresponding fit statistics as a measure of the model fit. Note that the two RT approaches and the baseline model had the same number of PCs based on 5% of the sample size of the country/economy. Lower values (i.e., higher absolute value) indicate a better fit when the same item responses and the same number of PCs were used (Educational Testing Service [ETS], 2012). Although the same number of PCs was retained across the three models for each country/economy, the composition of the PCs changed. Therefore, the comparison shows which set of PCs provide the best relative predictive power. As Table H.4 shows, for all 11 countries/economies, the RT-by-item-type-and-hour approach always showed better fit compared to both the baseline model and the RT-by-hour approach.

Table H.4. Comparison of model fit across three models

Country/Economy	Number of PCs	Baseline	RT-by-hour	RT-by-item-type-and-hour
AUS	714	-17631.4	-18559.8	-19093.5
BRA	538	-19912.3	-20273.3	-20565.7
CRI	330	-26374.7	-26496.1	-26664.0
DEU	267	-7461.33	-7621.59	-7835.68
FRA	315	-10213.1	-10400.2	-10645.9
GRC	320	-21281.6	-21478	-21671.2
JPN	305	-8609.26	-8824.76	-8966.47
KOR	332	-21372.1	-21598.6	-21817.9
MEX	365	-13705.4	-13815.8	-13908.4
NLD	195	-6158.64	-6364.11	-6436.61
USA	253	-7074.13	-7300.73	-7349.27

PV-based reliabilities and correlations among the 10 PVs were further calculated to evaluate the performance of the three models. Table H.5 reports the PV-based reliabilities (OECD, 2017) from the three different models, and Table H.6 presents the distribution of the correlations among the 10 PVs after aggregating across the 11 countries/economies. Both tables show that only negligible differences were observed with respect to these PV-based statistics across the three different models.

Table H.5. Comparison of PV-based reliabilities across three models

	Baseline			RT-by-hour			RT-by-item-type-and-hour		
	Math	Reading	Science	Math	Reading	Science	Math	Reading	Science
AUS	0.842	0.927	0.880	0.842	0.929	0.881	0.842	0.928	0.880
BRA	0.826	0.922	0.858	0.825	0.922	0.862	0.825	0.923	0.865
CRI	0.805	0.905	0.845	0.802	0.906	0.847	0.804	0.906	0.849
DEU	0.870	0.933	0.894	0.870	0.933	0.896	0.868	0.933	0.897
FRA	0.864	0.933	0.884	0.867	0.932	0.887	0.864	0.935	0.887
GRC	0.802	0.926	0.846	0.805	0.926	0.851	0.807	0.926	0.852
JPN	0.838	0.919	0.877	0.837	0.920	0.876	0.839	0.919	0.876
KOR	0.852	0.919	0.878	0.850	0.920	0.878	0.850	0.920	0.882
MEX	0.803	0.905	0.851	0.798	0.907	0.850	0.801	0.906	0.849
NLD	0.858	0.930	0.881	0.861	0.931	0.883	0.862	0.931	0.884
USA	0.865	0.937	0.890	0.864	0.936	0.893	0.864	0.937	0.892

Table H.6. Distribution of pairwise correlations among 10 PVs across three models

	Baseline		RT-by-hour		RT-by-item-type-and-hour	
	Mean	SD	Mean	SD	Mean	SD
Mathematics	0.880	0.001	0.880	0.001	0.880	0.001
Reading	0.935	0.000	0.935	0.000	0.935	0.000
Science	0.898	0.001	0.899	0.001	0.900	0.001

At the country/economy level, note that the actual differences across the three approaches in the residual variances, posterior means, and posterior SDs were very small. For instance, Table H.7 compares the residual variance estimates for the three approaches for mathematics, reading, and science across the 11 countries/economies. This table shows that the RT-by-item-type-and-hour approach always showed the lowest or almost the lowest residual variance estimates in all domains in all countries/economies, suggesting that the predictive power was increased in the LRM with PCs extracted under this approach. When the country/economy-level posterior means and SDs were examined, Tables H.8A and H.8B showed that the three approaches resulted in almost identical results. Thus, including RT would not threaten the trends.

Table H.7. Difference in residual variances for two alternative approaches against the baseline approach in mathematics, reading, and science (before transformation to the PISA scale)

	Baseline - By hour			Baseline - By item type and hour		
	Mathematics	Reading	Science	Mathematics	Reading	Science

AUS	0.007	0.016	0.007	0.012	0.023	0.012
BRA	0.002	0.005	0.003	0.005	0.009	0.005
CRI	0.001	0.001	0.001	0.002	0.004	0.002
DEU	0.003	0.006	0.003	0.007	0.014	0.007
FRA	0.002	0.004	0.003	0.005	0.011	0.006
GRC	0.003	0.004	0.003	0.007	0.010	0.005
JPN	0.005	0.007	0.003	0.009	0.010	0.006
KOR	0.003	0.006	0.004	0.008	0.011	0.008
MEX	0.001	0.002	0.001	0.001	0.005	0.002
NLD	0.005	0.007	0.005	0.006	0.010	0.006
USA	0.004	0.009	0.006	0.005	0.011	0.008

Tabe H.8A. Difference in posterior means for two alternative approaches against the baseline approach in mathematics, reading, and science (before transformation to the PISA scale)

	Baseline - By hour			Baseline - By item type and hour		
	Mathematics	Reading	Science	Mathematics	Reading	Science
AUS	0.000	0.001	0.000	0.000	0.002	0.002
BRA	-0.001	-0.001	-0.001	0.001	-0.001	-0.002
CRI	-0.002	-0.001	0.000	-0.002	-0.001	0.000
DEU	0.000	0.000	0.000	0.000	0.000	0.000
FRA	-0.002	0.000	0.002	-0.002	0.000	0.002
GRC	0.001	0.000	-0.002	0.000	0.000	-0.001
JPN	0.000	0.001	0.001	-0.003	0.000	0.002
KOR	0.000	0.000	0.001	0.000	0.000	0.002
MEX	0.000	0.000	-0.001	0.001	-0.001	0.000
NLD	0.000	0.000	0.002	0.001	0.000	0.001
USA	0.000	0.001	0.001	0.000	0.001	0.001

Table H.8B. Difference in posterior SDs for two alternative approaches against the baseline approach in mathematics, reading, and science (before transformation to the PISA scale)

	Baseline - By hour			Baseline - By item type and hour		
	Mathematics	Reading	Science	Mathematics	Reading	Science
AUS	0.000	0.001	0.001	0.001	0.001	0.002
BRA	-0.001	0.000	0.000	-0.001	0.000	0.001
CRI	0.000	0.000	0.001	0.002	0.000	0.001
DEU	0.000	0.000	0.001	0.000	0.000	0.001
FRA	-0.001	0.000	0.001	0.002	0.000	0.001
GRC	-0.001	0.000	0.000	-0.002	0.000	0.000
JPN	-0.001	0.000	0.001	0.001	0.000	0.001
KOR	0.000	0.000	0.001	0.000	0.001	0.000

MEX	0.000	0.000	0.001	0.001	0.000	0.001
NLD	0.002	0.001	-0.001	0.000	0.001	-0.001
USA	-0.001	0.001	0.002	-0.002	0.000	0.002

Although PISA is not targeted at individual-level reporting, a further comparison at the individual level was conducted across the three different approaches. As an example, the posterior means and posterior SDs for the three PISA domains of one country/economy are presented in Figures H.2A and H.2B. Figure H.2A indicates that the baseline model and the RT-by-item-type-and-hour approach generated similar individual-level posterior means, and it appears the results are more similar between the two models for reading than for mathematics or science. This is probably because, according to the design, all students took reading, while some students did not take mathematics or science, thus, their results are purely model dependent. Figure H.2B shows the differences in individual-level posterior SDs between the baseline model and the RT-by-item-type-and-hour approach. For reading and science, most of the students had smaller posterior SDs at the individual level when the RT-by-item-type-and-hour approach was used, indicating that measurement precision improved with this approach. For mathematics, at least half the students showed improved measurement precision when the RT-by-item-type-and-hour approach was used. Although not presented, this pattern of reduced posterior SD at the individual level under the RT-by-item-type-and-hour approach was consistently observed for all countries/economies.

Figure H.2A. Comparison of individual-level posterior means between the baseline model and the RT-by-item-type-and-hour approach for mathematics, reading, and science (after transformation to the PISA scale) for a certain country/economy

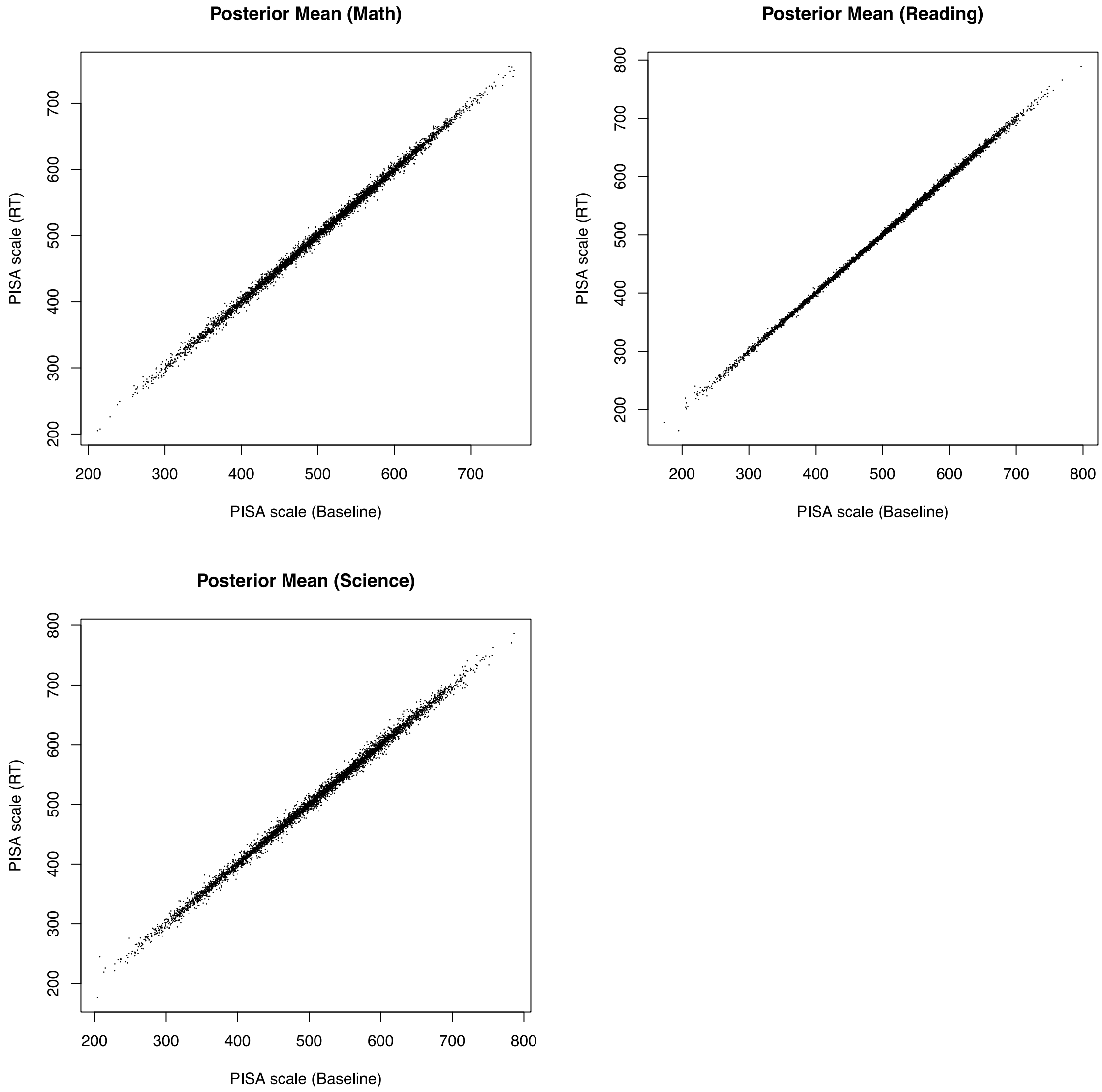
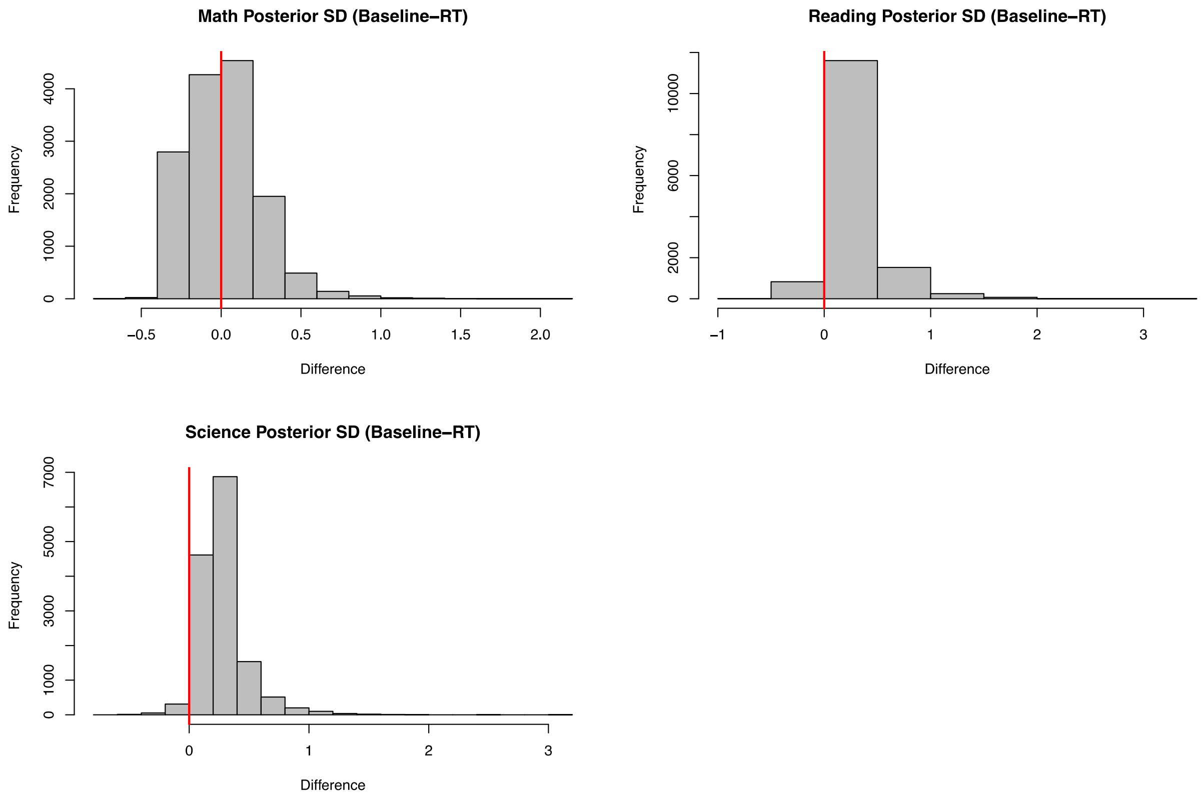


Figure H.2B. Histogram of differences in individual-level posterior SDs between the baseline model and the RT-by-item-type-and-hour approach for mathematics, reading, and science (before transformation to the PISA scale) for a certain country/economy



This set of experimental studies can be summarised as follows. First, a similarity in country/economy- and individual-level posterior means was observed across the three approaches, implying that there would not be much impact if RT information was used in the population modelling for the 2018 cycle. Second, given that there is no expected harm to the trends, smaller residual variance estimates and better model fits in all 11 countries/economies suggest that including RT information can improve the predictive power in the LRMs. Third, such gains seem to have been obtained through improved measurement precision at the individual level with reduced posterior SDs. Therefore, RT information was incorporated in the PISA 2018 population modelling using the RT-by-item-type-and-hour approach as described above. Further studies will be conducted to optimise the process of extracting information from the RT variables without introducing biases through the violation of the conditional independence assumption of the IRT-LRMs.

Change 3: Modelling the Financial Literacy Sample

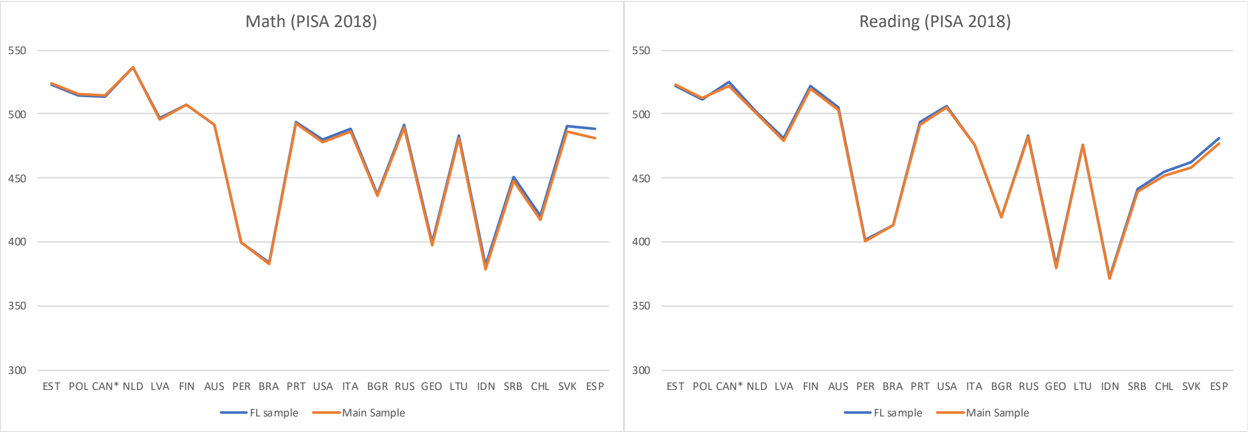
The assessment of financial literacy was offered as an international option in PISA 2018. In total, 21 countries/economies opted to administer this assessment. The cognitive instruments included trend items from 2012 and 2015, as well as a set of new interactive items that were developed specifically for PISA 2018. Note that financial literacy was available only in the computer-based assessment mode. As a reminder, in PISA 2015, the financial literacy assessment was administered to a subset of students from the main sample during additional testing time. However, in 2018, financial literacy was administered to a separate sample of PISA-eligible students who took a combination of reading, mathematics, and financial literacy items. The group of students who took the financial literacy assessment are referred to as the *Financial Literacy sample*.

Countries/economies administering the financial literacy instruments required 1,650 additional students in the sample. Each student taking the financial literacy assessment took the financial literacy items in addition to the mathematics or reading items, with the reading items administered in the same adaptive mode as in the main sample, including the reading fluency tasks. Students taking the financial literacy assessment did not take any science or global competence items. Therefore, financial literacy sample students received PVs in mathematics, reading, and financial literacy.

The financial literacy sample data were used for the IRT scaling to estimate the financial literacy item parameters, but were not used to estimate the reading or mathematics item parameters. When estimating the multivariate latent regression models, the financial literacy sample was combined with the sample of students from the main sample who took reading and mathemetics only. This was done to establish a stable linkage between the financial literacy and the main PISA forms and between the reading and mathematics domains. RT information was used as conditioning variables in all countries/economies, but the newly developed school-level WLEs as continuous covariates was applied only to the four countries/economies with a significant amount of oversampling (i.e., Indonesia, Italy, Russian Federation, and Spain).

By design, the financial literacy sample is a randomly selected group of students, and thus, their posterior distributions are expected to be comparable to those of the main sample students in mathematics and reading. Figure H.3 presents the comparison of the posterior means in mathematics and reading between the main sample and the financial literacy sample for the 21 countries/economies that administered financial literacy. Except for one country/economy (which had a very large amount of cases in the main sample and only a small amount of cases in the financial literacy sample), all countries/economies showed almost identical posterior means in reading and mathematics.

Figure H.3. Posterior means of the main sample and the financial literacy sample for mathematics and reading (after transformation to the PISA scale)



Change 4: Modelling Large Proportions of UH Cases

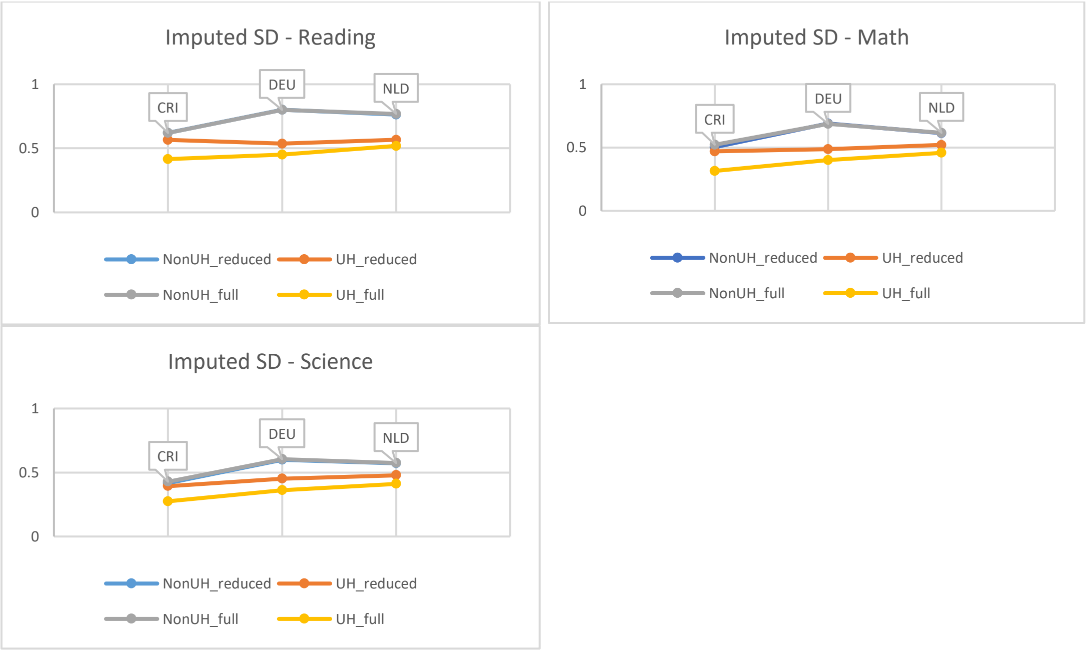
The UH assessment option for special-needs students was provided in 2015 and 2018. Students who took the UH assessment are different in their representativeness of the target population and in the instruments that were administered to them, both for the cognitive assessment and the BQ. Students taking the UH instrument take the UH version of the student BQ, which consists of a subset of items from the regular student BQ. Due to these differences, in this experiment, the proficiency estimates of the UH students were calculated through a reduced conditioning model that involved only the subset of BQ variables included in the UH version.

The experiment here was designed to see if the reduced conditioning model would generate results comparable to the full model that used all items in the regular BQ. Instead of including the UH cases in a single conditioning model estimated in each country/economy, as was done in 2015, a mixed approach was attempted. In this mixed approach, to generate PVs for the non-UH cases, the population model parameters were estimated using only the non-UH cases and the entire set of BQ questions (full model). In contrast, to generate PVs for the UH cases, the population model parameters were estimated using the entire sample (including both the UH cases and non-UH cases), but only the subset of BQ questions administered to the UH students was included (reduced model). The full model is used to generate PVs for the non-UH students, and the reduced model is used to generate PVs for UH students – PVs for non-UH cases generated from the reduced model were discarded. For the reduced model, the UH indicator variable was specified as a direct dummy-coded covariate instead of being processed through the PCA. For the full model, the UH indicator variable was not used because the full model only involved non-UH cases.

For the experiment, data from Costa Rica, Germany, and the Netherlands were used, as these three countries/economies were considered to have a relatively large number of UH students.[10] At the overall country/economy level, the full model and the reduced model generated similar results in terms of residual variances, cross-subject correlations, posterior means, and posterior SDs for mathematics, reading, and science. As expected, the full model generated slightly smaller residual variances for all three domains. The dataset of each country/economy was then disaggregated into UH and non-UH cases, and the proficiency estimates for the two models were compared. For the non-UH cases, the group-level posterior means and SDs for the two models were similar. For the UH cases, the posterior mean estimates were similar between the two models, but the posterior SD estimates were smaller for the full model, as shown in Figure H.4. This was interpreted as a sign of overfitting, since the additional variables included in the full model did not contain any information about the UH cases. The comparisons suggested that the reduced model generated reasonable proficiency estimates for the UH cases. Therefore, a mixed approach was used to produce PVs for the five countries/economies in which more than 200 cases included in the main sample were UH cases (i.e., Canada-English speaking, Canada-French speaking, Costa Rica, Denmark, and the Netherlands). The 2015 approach was used for the other countries with a smaller number of UH cases.

[10] In 2018, Costa Rica, Germany, and the Netherlands had 607 UH cases, 98 UH cases, and 851 UH cases, respectively. Given that the new mixed approach was applied to countries/economies in which more than 200 cases included in the main sample were UH cases, the new approach was not used to generate PVs for UH cases in Germany, even though Germany had been included in the experiment.

Figure H.4. Comparison of posterior SD estimates for UH and non-UH cases between the full and reduced models for mathematics, reading, and science (before transformation to the PISA scale)



REFERENCES

Benton, T. (2019). The effect of using principal components to create plausible values. In M. Wiberg, S. Culpepper, R. Janssen, J. González, & D. Molenaar (Eds.), *Quantitative psychology* (pp. 293–306). Springer. https://doi.org/10.1007/978-3-030-01310-3_26

Educational Testing Service. (2012). *DGROUP* [Computer software].

Glas, C. A. W. (1988). The Rasch model and multistage testing. *Journal of Educational Statistics*, *13*(1), 45–52. https://doi.org/10.3102/10769986013001045

Goldhammer, F., Martens, T., Christoph, G., & Lüdtke, O. (2016). *Test-taking engagement in PIAAC* (OECD Education Working Papers No. 133). OECD Publishing. https://doi.org/10.1787/5jlzfl6fhxs2-en

Jewsbury, P., Lu, R., & van Rijn, P. W. (2019). *Modeling multistage and targeted testing data with item response theory* [Manuscript submitted for publication]. Research and Development Division, Educational Testing Service.

Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 31*(3), 300–303. https://doi.org/10.2307/2348005

Lee, Y.-H., & Haberman, S. (2016). Investigating test-taking behaviors using timing and process data. *International Journal of Testing*, *16*(3), 240–267. https://doi.org/10.1080/15305058.2015.1085385

Lee, Y.-H., & Jia, Y. (2014). Using response time to investigate students' test-taking behaviors in a NAEP computer-based study. *Large-Scale Assessments in Education*, *2*(8). https://doi.org/10.1186/s40536-014-0008-1

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input. *Statistical Science, 9*(4), 538–558. https://www.jstor.org/stable/2246252

Meyer, J. P. (2010). A mixture Rasch model with item response time components. *Applied Psychological Measurement*, *34*(7), 521–538. https://doi.org/10.1177/0146621609355451

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*(2), 177–196. https://doi.org/10.1007/BF02294457

Organisation for Economic Co-Operation and Development. (2013). *Technical report of the Survey of Adult Skills (PIAAC).* https://www.oecd.org/skills/piaac/Skills_Matter_Further_Results_from_the_Survey_of_Adult_Skills.pdf

Organisation for Economic Co-Operation and Development. (2017). *PISA 2015 technical report.* https://www.oecd.org/pisa/data/2015-technical-report/

Shin, H. J., Jewsbury, P., & van Rijn, P. W. (2019, October). *Conditional dependencies between cognitive responses and process data in large-scale assessments* [Poster presentation]. Foundational and Applied Statistics and Psychometrics poster session, Educational Testing Service, Princeton, NJ, United States.

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

Shin, H. J., Yamamoto, K., Khorramdel, L., Robin, F., von Davier, M., Gamble, H., & Zhao, W. (2019, April). *Incorporating response time into the PISA population model* [Paper presentation]. National Council on Measurement in Education (NCME) Annual Meeting, Toronto, Canada.

Thomas, N. (2002). The role of secondary covariates when estimating latent trait population distributions. *Psychometrika*, *67*(1), 33–48. https://doi.org/10.1007/BF02294708

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365–384. https://doi.org/10.1007/s11336-007-9046-8

van der Linden, W. J., & Sotaridona, L. (2006). Detecting answer copying when the regular response process follows a known response model. *Journal of Educational and Behavioral Statistics*, *31*(3), 283–304. https://doi.org/10.3102%2F10769986031003283

van Rijn, P. W., & Shin, H. J. (2019). Item calibration for multistage tests in the context of large-scale educational assessment [Manuscript in preparation]. Research and Development Division, Educational Testing Service.

von Davier., M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2014). 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%2F1076998619881789

von Davier, M., & Sinharay, S. (2014). Analytics in international large-scale assessments: Item response theory and population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 155–174). CRC Press.

Weeks, J., von Davier, M., & Yamamoto, K. (2016). Using response time data to inform the coding of omitted responses. *Psychological Test and Assessment Modeling*, *58*(4), 671–701.

Yamamoto, K. (2019, May). *Using timing information associated with response data in large-scale assessment* [Paper presentation]. The Opportunity versus Challenge: Exploring Usage of Log-File and Process Data in International Large-Scale Assessments Conference and Workshop, Dublin, Ireland.

Yamamoto, K., Shin, H. J., & Khorramdel, L. (2019), *Introduction of multistage adaptive testing design in PISA 2018* (OECD Education Working Papers No. 209). OECD Publishing. https://doi.org/10.1787/b9435d4b-en