Annex K: Uses and Reporting of Process Data

In this annex, the general topic of process data and defining process variables in PISA is discussed in Section 1. In Section 2, results with respect to the recently discovered issue with response time variables in case of to multiple item visits in PISA 2015 and 2018 are presented.

1. Process Data in PISA

In PISA, the switch to computer-based assessment (CBA) as the main mode of administration in 2015 has opened up possibilities to gather additional information on how students interact with the items. This additional information is collectively referred to as process data and can be defined as any data about the response process. In general, such data can vary from relatively basic, such as data related to the timing of responses, to relatively sophisticated, such as eye-tracking data of students during the assessment. Process data has two types of potential uses: 1) enhancing data quality and security, and 2) improving reliability and validity (Ercikan & Pellegrino, 2017). In PISA, process data are captured by writing test events to a log file for each student's session. It is collected for both the cognitive assessment and the background questionnaire (BQ), but the focus here is on process data from the cognitive assessments.

With respect to the goals of PISA, how process data can inform group comparisons and how they relate to the proficiency distributions may be most relevant (Ercikan et al., 2020). Although some process data are available from earlier cycles, since 2015, a number of process variables have been added to the public user files (PUF). These are response time, number of actions, and, since 2018, response time to first action. The availability of PISA process data has led to considerable research on a variety of topics: The treatment of missing data (Lu & Wang, 2020; Ulitzsch et al., 2019a), problem-solving strategies (Greiff et al., 2015; He et al., 2018), student effort and engagement (Anaya & Zamarro, 2020; Michaelides et al., 2020; Ulitzsch et al., 2019b), response-time scaling (Shin et al., 2020), and profiles of student inquiry (Teig et al., 2020). However, most research has been conducted on relatively small subsets of items (von Davier et al., 2019a).

1.1 Logged Events in PISA

Process data are captured by writing timestamped test events to a log file for each student's session. Table 1 below shows all the different logged events in the student delivery system (SDS) in PISA. Not all events or event details are logged in PISA, and the logged events vary by item type and presentation on the computer screen. For example, keypresses are logged, but not which keys. This prevents the differentiation of types of response behaviors (e.g., text production vs. editing). Keystroke logs could be interesting for detecting group differences (e.g., Zhang et al., 2019). In addition, two items sometimes appear on one screen, so that the process data cannot always be clearly connected with

an individual item. It also needs to be mentioned that the log files can be large and noisy. For example, there are cases in PISA were test administrators needed to move students to a different computer due to technical problems and would either restart the test entirely or fast-forward them to the place where they left off. Fixing and merging log files in these cases is not straightforward.

Table 1 Alphabetic List of Logged Events in PISA SDS

1. breakoff	12. paste
2. change	13. PSReadyToScore
3. click	14. PSscored
4. copy	15. questionLoaded
5. cut	16. run_simulation
6. dblclick	17. scoreNowResult
7. drop	18. scoring
8. keypress	19. selectionevent
9. move	20. selectWord
10. onItemBegin	21. stimulusAndQuestionLoaded
11. onItemEnd	22. stimulusLoaded

1.2 Defining Process Variables

For PISA, a distinction can be made between the process of *creating* new variables from process data and *publishing* new process variables in the PUF. Creating new variables from process data is complex, and, across the field of educational measurement, this complexity is only beginning to be understood. While there are plenty of guidelines for quality assurance of raw and scored item responses (e.g., item analysis, distractor analysis, rater agreement), the literature on standards of practice for creating process variables and their quality assurance is limited. With its growing expertise in this area, ETS has determined that creating new process variables requires a solid QC process with the elements listed below:

1. The computer code that produces the variable + documentation (e.g., Javascript, Python, SAS, R)
2. Properties of new process variables (e.g., mean, SD, distribution)
3. Relations with existing variables (e.g., crosstables, correlations, scatter plots)

Process variables are derived from the event data in the log. Derivation of the variable is complex, involving an exploration of the event data to define the variable and the development and checking of the computer code to calculate the variable for each test log. Errors can arise in the logic used to define the variable or the code written to calculate it. Each of the above steps needs review to ensure data accuracy. The quality control for each process variable should also include the collection of evidence to support the validity of the meaning assigned to the derived variable.

2. Updated Timing Variables in PISA 2015 and 2018

This section turns from a general discussion of process variables to a review of a specific process variable: time spent on an item. In 2020, it was found that the item-level response time (RT) variable

T, as reported for the cognitive items in the PUF for PISA 2015 and 2018, was the time spent on the last visit to an item. However, since students are able to move back and forth among items within a unit, this variable does not always equal the total time spent on an item. Total time spent is the timing information of greatest interest to researchers who will use the PUF. The difference between time spent on last visit and total time has an impact on reported results with respect to timing. Furthermore, derived timing variables were used in the conditioning model in PISA 2018, and, therefore, this could have impacted the reported plausible values.

To illustrate that the definition of RT is not straightforward, the next section discusses how response time data is defined. Then, the impact of multiple item visits on reported summary statistics of timing variables is discussed. Since the re-analysis of the 2015 data is still ongoing at the time of this writing, results for 2018 are presented only. In addition, the impact on reported plausible values in PISA 2018 is discussed.

2.1 Defining Response Time Variables

When investigating the reported RT variables in the 2015 and 2018 PUF, it was found that defining a response time variable at the item level is not necessarily straightforward, and there are multiple ways to do it. For example, recalling Table 1, RT can be defined as the difference between timestamps of an item beginning and ending event: T(onItemEnd) – T(onItemBegin). However, three events typically occur after onItemBegin: stimulusLoaded, questionLoaded, and stimulusAndQuestionLoaded. It is reasonable to say that students can only start working on the item once these are done. These events create differential overhead depending on the computer, the size of stimulus and question, and the number of item visits. An alternative definition would be to use the difference between timestamps of the last item loading event and an item ending: T(onItemEnd) – T(stimulusAndQuestionLoaded). However, the student delivery system (SDS) in 2015 and 2018 allowed students to click on next before stimulusAndQuestionLoaded, which can result in negative RT.

Based on these differences, a distinction was made between the duration of an item and the item-level RT (reported as *TT* or *total time* in the additional PUF[1]). The *duration* includes all the overhead time, including introduction screens, loading events, and (automated) scoring events. The RT variable *TT* does not include this overhead. With students being able to visit items multiple times within a unit, multiple timing variables can be defined. In addition, process variables related to item visits, actions, and scoring can be defined. The list of variables is shown in Figure 1, and the highlighted variables are those included in the additional PUF. Except duration of unit, all variables listed are defined at the item level. With these variables, several crosschecks can be defined for quality control purposes (e.g., sum

[1] This additional PUF, *cognitive items total time/visits data file*, was uploaded on the PISA website in September 2020 for PISA 2018 and on (http://www.oecd.org/pisa/data/2018database/) and in November 2020 for PISA 2015 (http://www.oecd.org/pisa/data/2015database/).

of item duration is smaller than or equal to unit duration). In addition, these variables may be aggregated to provide useful summaries (e.g., total number of revisits, score changes).

Figure 1 List of new process variables (highlighted variables are part of the additional PUF)

- Timing:
    1. Duration of unit
    2. Duration across all item visits
    3. Duration of first item visit
    4. Response time across all item visits (*TT*)
    5. Response time on first item visit
    6. Response time to first action
- Visits:
    7. Number of item visits (*V*)
    8. Number of short item visits (< .5s)
- Actions:
    9. Number of actions across all item visits
    10. Number of actions on first item visit
- Scoring
    11. Score change across multiple item visits

2.2 Impact on Timing Results in PISA 2018

Since the timing issue is related to multiple item visits, this section illustrates the extent to which multiple item visits were observed. Figure 2 below shows the log frequency of the number of item visits across math, reading, science, and global competence across all participants (n = 551,930). Even after taking the logarithm, there seems to be an exponential decrease in the frequency with increasing item visits.

Figure 2 Log frequency of number of item visits across math, reading, and science in 2018 (n = 551,930)

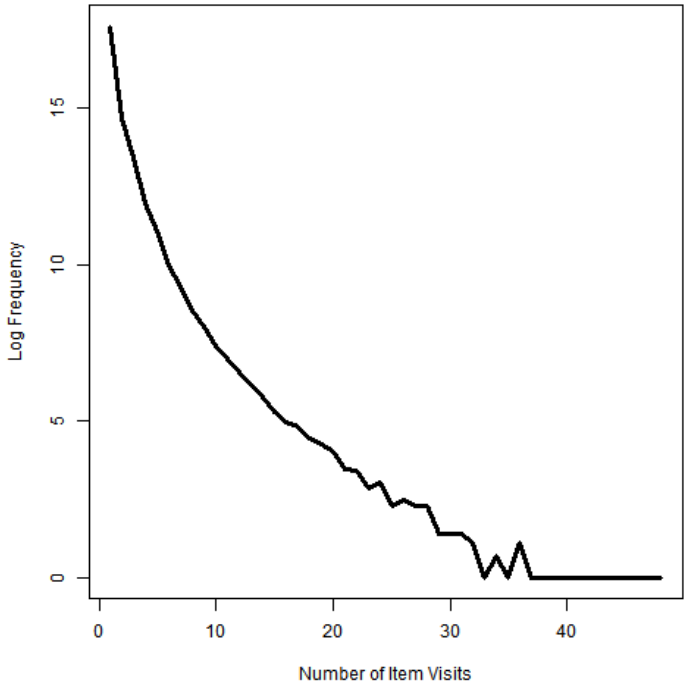


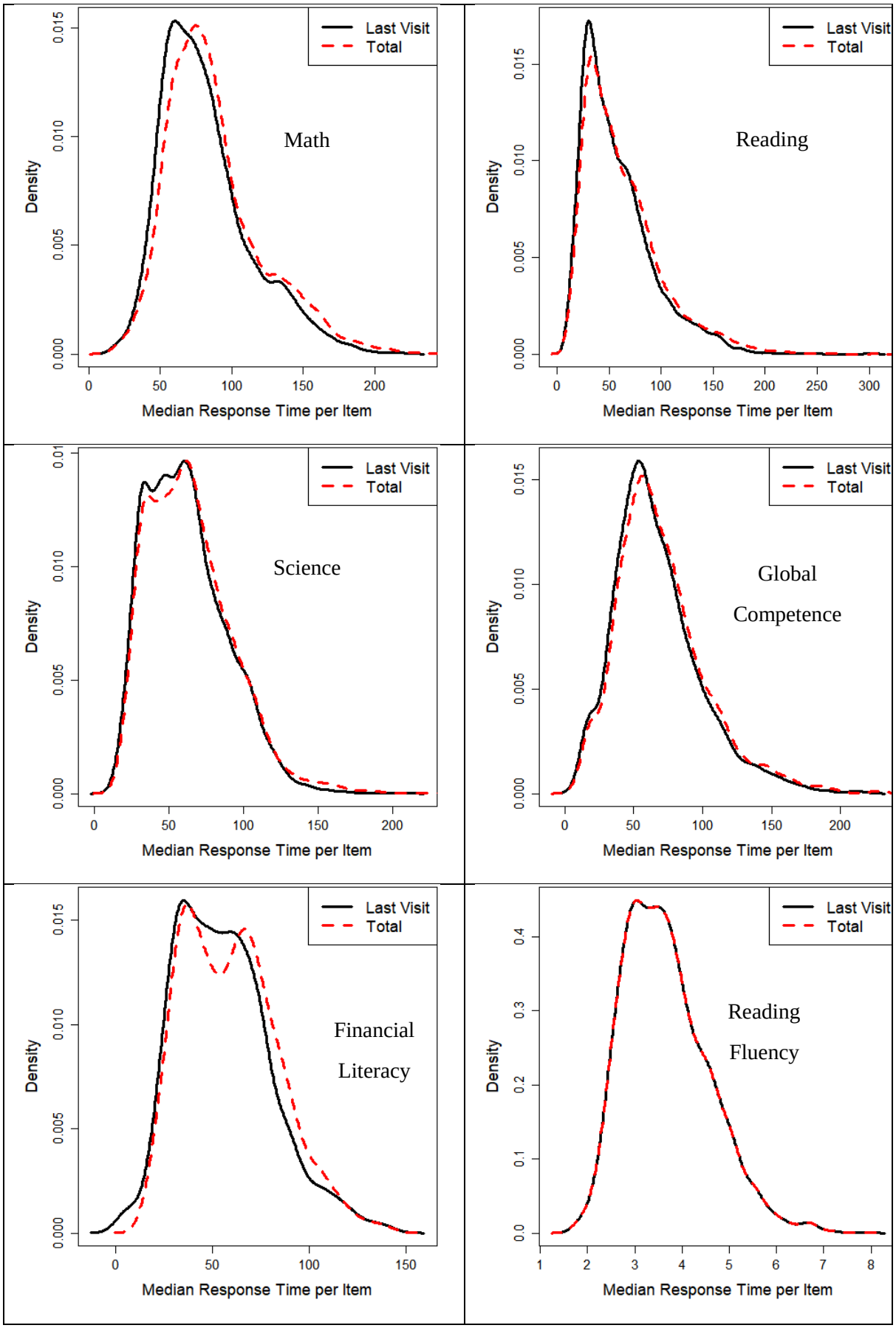
Table 2 shows the percentages of the frequency distribution of the number of item visits in 2018. The data comprises 551930 students and 47278588 item visits. Overall, in 93% of the cases, items were visited only once, and the originally published RT variable, time on last visit, or *T*, is not affected. However, in 7% of the cases, items were visited multiple times, and the RT variable *T* is affected.

Table 2 Distribution of Number of Item Visits in 2018

Number of Item Visits	Percentage	Cumulative Percentage
1	93.0	93.0
2	5.0	98.0
3	1.5	99.4
4	0.3	99.7
5	0.1	99.9
More than 5	0.1	100.0

Figure 3 shows the distributions of the medians of the two RT variables per item for each participating country/economy across domains. The solid black line shows the distribution using the old definition (time on last visit, *T* in the original PUF), and the dashed red line shows the distribution for the total time across all visits using the new definition (*TT* in the additional PUF). It can be seen that with the new definition, median RT is larger across domains with the exception of reading fluency. This is because the SDS for reading fluency did not allow students to go back to previous items.

Figure 3 Distributions of median response time per item for different domains across all participants



Based on the figures above, the overall impact does not appear to be large. However, when checking the median RT for different numbers of visits, more substantial differences are exposed. As shown in Table 3, the difference between median RT based on last visit (*T*) and median RT based on total (*TT*) is very large in the case of more than one visit (21s vs. 94s), even though this difference is seen for only 7% of responses. In addition, there is a noticeable change in the biserial correlation of item score and log RT in case of multiple item visits between RT based on the last visit and total (-.05 vs. 02). These results are aggregated across countries and items, so it cannot be ruled out that other differences in magnitude and direction may be found at different levels of analysis.

Table 3. Median RT and Biserial Correlation Between Item Score and Log RT Based on Last Visit and Total for Single and Multiple Item Visits for Math, Reading, and Science in PISA 2018

		Median RT		Biserial Corr.	
Item Visits	Percentage	Last Visit	Total	Last Visit	Total
Single	93.0	48s	48s	.13	.14
Multiple	7.0	21s	94s	-.05	.02
Overall	100.00	46s	51s	.08	.13

2.3 Impact on Plausible Values in PISA 2018

In PISA 2018, aggregated timing variables in the form of person-level RT deciles were incorporated in the population modelling in generating plausible values[2]. Hence, it is important to evaluate the impact of including the new timing variables in the generation of model-based plausible values, and the inferences on group statistics. This section describes the methods used to investigate the impact and the results. More details about how the RT data were processed and how they were included in the population modelling can be found in Annex H of the PISA 2018 technical report.

First, four countries were identified based on the difference in values between the timing variables *T* (last visit) and *TT* (total). Three countries/economies with substantial differences were identified and one country with relatively small differences was selected as a reference. Second, it was found that the procedure for computing the RT deciles had to be updated, as it did not fully incorporate the multistage adaptive testing (MSAT) design for reading.

These changes resulted in a set of updated PVs based on updated timing variables with *TT* replacing *T* and an updated procedure for deriving the person-level RT deciles. These updated PVs were carefully compared against the published PVs for the four selected countries/economies and selected subgroups.

[2] Those derived person-level RT decile variables were not included in the 2018 PUF. In PISA 2015, response time variables were not modeled in the population modeling, so "impact on plausible values" is only relevant for PISA 2018.

2.3.1 Methods

To identify countries for further investigation, the following variables were used:

- Item-level RT (T and TT);
- Number of item visits;
- Number of rapid responses;
- Person-level RT deciles.

In the comparison of PVs of selected countries, the following statistics were compared:

- Country-level summary statistics;
- Within-country subgroup summary statistics;
- Correlations among published and updated PVs.

On a theoretical basis, country-level summary statistics (e.g., mean and standard deviation) are not expected to change with the updated timing variables, because these are essentially determined by the international scaling model, not the country-level conditioning model. That is, in the item-response theory (IRT) model used to create the international PISA scale, the first three moments of the latent distribution are estimated for each country (e.g., von Davier et al., 2019b). This ability variable is then regressed on the conditioning variables. However, if subgroups differ in how they relate to the updated timing variables (e.g., females tend to go back more often to previous items than males), then differences in subgroup statistics may occur.

The difference between published PVs and updated PVs can be defined for each student i and PV u by

$$D_{iu} = PV_{iu,pub} - PV_{iu,upd},$$

The mean difference across multiple imputations is then

$$\overline{D} = \frac{1}{U}\sum_{i=1}^{U}\overline{D}_u,$$

where $\overline{D}_u = \frac{\sum_{i=1}^{N} w_i D_{iu}}{\sum_{i=1}^{N} w_i}$, the weighted mean difference for each pair of PVs and U, the number of imputations, equals 10. Generally, there are two sources of variance in the results: sampling variance and imputation (or measurement error) variance. If the differences in group statistics between published and updated PVs are within the margin of error due to sampling and imputation, then it can be argued that there is no need to update the published PVs. A confidence interval (CI) for the mean difference that reflects the sampling or imputation error can be constructed as

$$CI = \overline{D} \pm Z_{1-\frac{\alpha}{2}} \times S(\overline{D}),$$

where $Z_{1-\frac{\alpha}{2}}$ is the critical value associated with the nominal level α for the standard normal distribution, and $S(\overline{D}) = \sqrt{V(\overline{D})}$ is the standard error due to sampling or imputation associated with the difference in PVs. If the confidence interval does not contain zero, then the difference can be considered outside the margin of error. In evaluating the differences, 95% CIs are used in the present analyses, and separate CIs are constructed for sampling and imputation errors. However, if many differences are tested in this way, the problem of multiple comparisons occurs, and corrections to control the false discovery rate may need to be considered.

Within each of the four countries/economies, the following grouping variables were used to study the differences in published and updated PVs for subgroups: highest education of parents (HISCED), gender (ST004D01T), number of books at home (ST013Q01TA), math imputed, and science imputed. These last two variables are used to indicate students who did not take math or science by design in PISA 2018.

Finally, checks on the relation of the TT-based deciles with the published and updated PVs were performed. This is done because the updated deciles are uncongenial to the published PVs (Meng, 1994), and this may cause different relationships if TT variables are published files but the PVs are not updated accordingly. The following eight person-level RT deciles were used in the conditioning model: Average and standard deviation (SD) of RT in the first hour for human-coded items (AVE_1_HUM, STD_1_HUM), average and SD of RT in the first hour for machine-coded items (AVE_1_MAC, STD_1_MAC), average and SD of RT in the second hour for human-coded items (AVE_2_HUM, STD_2_HUM), and average and SD of RT in the second hour for machine-coded items (AVE_2_MAC, STD_2_MAC). In the comparisons, the focus is only on the averages.

2.3.2 Results

Table 4 below shows the results for published and updated PV-based means, standard deviations, and standard errors of the statistics for the four selected countries/economies—three (A, B and C) with substantial differences in values between the timing variables T (last visit) and TT (total) and one (D) with relatively small differences. The differences are generally small, with the largest difference seen in B for mathematics (423.15 vs. 421.70). However, none of the differences in means are statistically significant.

Table 4 Published and Updated PV-based Means, SDs, and SEs for Selected Countries/economies

		Math		Reading		Science	
Country/economy	Statistic	Published	Updated	Published	Updated	Published	Updated
A	Mean	591.39	591.44	555.24	555.17	590.45	591.00
	SD	80.33	80.26	87.23	86.96	83.19	83.46
	SE	2.52	2.68	2.75	2.70	2.67	2.67
B	Mean	423.15	421.70	386.91	386.93	397.10	397.25
	SD	86.96	87.31	77.33	77.39	75.72	75.38
	SE	1.91	2.03	1.46	1.45	1.66	1.63
C	Mean	453.51	454.15	465.63	465.66	468.30	468.11
	SD	88.16	88.72	87.66	87.92	83.53	82.87
	SE	2.26	2.38	2.17	2.16	2.01	2.22
D	Mean	500.04	499.87	498.28	497.93	502.99	502.52
	SD	95.39	95.34	105.75	105.79	102.86	102.58
	SE	2.65	2.78	3.03	3.01	2.91	2.91

Figure 4 below shows the percentiles (5%,25%,50%,75%, and 95%) of the distributions of the published and updated PVs for the selected four countries/economies on the three core domains. The size of the blocks indicates the 95% CI around the percentile. The percentile distributions generally match between published and updated PVs and differences tend to be small and within the standard errors.

Figure 4 Percentiles of published and updated PV distributions of selected countries/economies

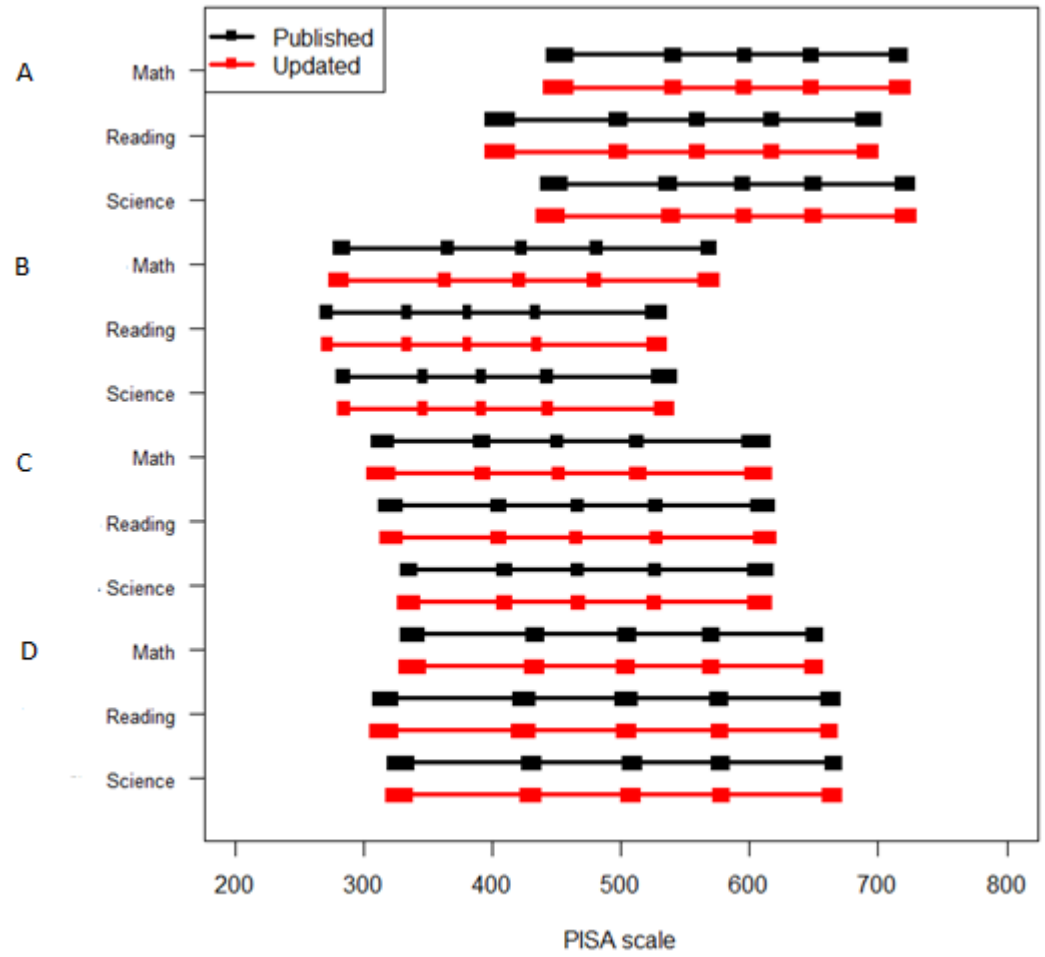


Table 5 below shows the average correlations within published PVs, within updated PVs, and between published and updated PVs. The average correlations within PVs can be interpreted as reliabilities. The largest differences in correlations are seen in mathematics for B. Generally, the differences are small (third decimal), especially for the major domain reading. The differences for the reference country D are smallest.

Table 5 Average Correlations of Published and Updated PV for Selected Countries/economies

Country/economy	Correlation	Math	Reading	Science
A	Within Published PV	0.837	0.908	0.870
	Within Updated PV	0.841	0.906	0.870
	Between Published and Updated PV	0.829	0.906	0.865
B	Within Published PV	0.762	0.928	0.855
	Within Updated PV	0.771	0.927	0.852
	Between Published and Updated PV	0.753	0.925	0.843
C	Within Published PV	0.851	0.924	0.877
	Within Updated PV	0.854	0.921	0.877
	Between Published and Updated PV	0.844	0.921	0.871
D	Within Published PV	0.878	0.936	0.902
	Within Updated PV	0.877	0.936	0.901
	Between Published and Updated PV	0.876	0.936	0.899

Finally, Table 6 shows the summary statistics for one set of reading subscales (evaluating and reflecting, locating information, and understanding). Again, differences are generally small (<0.5).

Table 6 Summary Statistics for Reading Subscale for Selected Countries/economies

		Evaluate		Locate		Understand	
Country/economy	Statistic	Published	Updated	Published	Updated	Published	Updated
A	Mean	565.12	564.95	552.66	552.26	561.9	561.9
	SD	93.01	92.46	92.96	92.12	86.53	86.49
B	Mean	388.81	388.59	389.47	389.92	394.2	394.4
	SD	83.09	82.93	82.67	82.82	75.17	74.97
C	Mean	474.57	473.97	462.7	463.2	473.77	474.26
	SD	95.88	96.18	89.08	88.79	88.03	88.15
D	Mean	496.69	496.74	497.97	498.03	494.32	494.45
	SD	109.93	110.14	113.15	113.05	108.47	108.9

Figure 5 displays the 95% confidence intervals for the differences in the subgroup statistics between the published and updated PVs, defined by the five grouping variables: highest education of parents (HISCED), gender (ST004D01T), number of books in the home (ST013Q01TA), math imputed, and science imputed. The left panels show the confidence intervals computed with the sampling error of the differences and the right panels show the confidence intervals computed with the imputation error of the differences. The sampling errors of the differences are generally smaller than the imputation errors of the differences. In both cases, however, only one out of the 228 confidence intervals flags a significant result: For the group of students in B with 11-25 books, the difference in math between published and updated PV is significant.

Figure 5 95% CIs for differences in subgroup statistics between published and updated PVs (left: with sampling error, right: with imputation error)

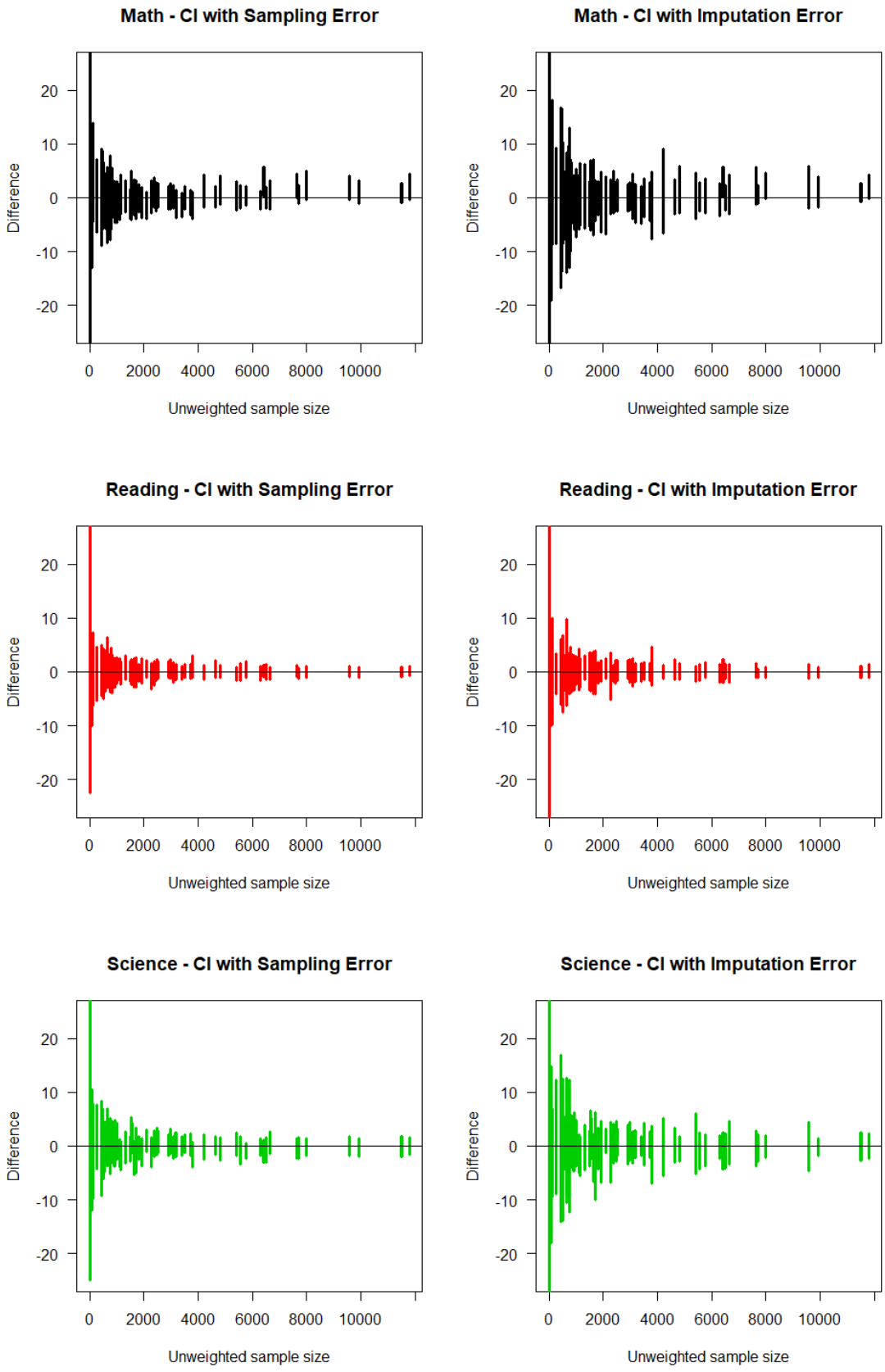
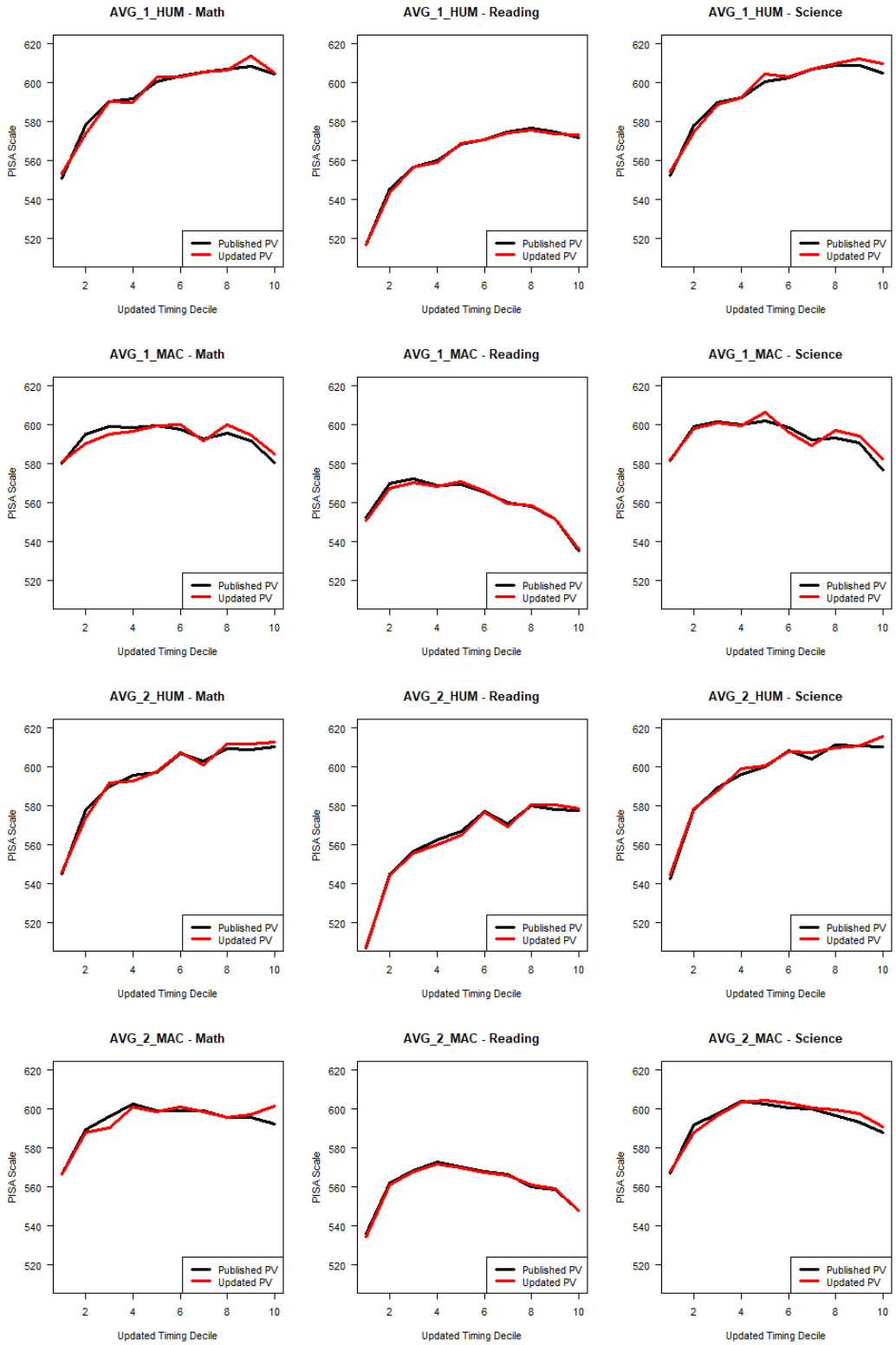


Figure 6 show the relation between the updated person-level RT deciles and the published and updated PVs for a selected country/economy A. Even though the timing variables for this country were most affected, the pattern of relations with the PVs does not change drastically.

Figure 6 Relation between updated RT deciles and published and updated PVs for a selected country/economy A



Based on the findings of analyses with the newly calculated timing variables (*TT*, total time across all item visits) for a selection of countries from PISA 2018, it can be concluded that the impact of incorporating derived RT variables in the models on the generation of the plausible values is limited. Although some differences were seen in some statistics, these differences were generally within the margin of sampling and imputation errors of the published results. Given that our selection consists of countries/economies with large differences in timing variables and a reference country with small differences, it is expected that the results for other countries/economies will fall within the range of results found here.

REFERENCES

Anaya, L., & Zamarro, G. (2020). *The role of student effort on performance in PISA: Revisiting the gender gap in achievement.* Education Reform Faculty and Graduate Students Publications. Retrieved from https://scholarworks.uark.edu/edrepub/116

Ercikan, K., Guo, H. & He, Q. (2020): Use of response process data to inform group comparisons and fairness research. *Educational Assessment.*

Ercikan, K., & Pellegrino, J. W. (Eds.). (2017). *Validation of score meaning for the next generation of assessments: The use of response processes*. Taylor & Francis.

Greiff, S., Wüstenberg, S., & Avvisati, F. (2015). Computer-generated log-file analyses as a window into students' minds? A showcase study based on the PISA 2012 assessment of problem solving. *Computers & Education, 91*, 92-105.

He, Q., von Davier, M., & Han, Z. (2018). Exploring process data in problem-solving items in computer-based large-scale assessments. In H. Jiao, R. W. Lissitz, & A. Van Wie (Eds.), *Data analytics and psychometrics: Informing assessment practices* (pp. 53-76). Information Age Publishing.

Lu, J., & Wang, C. (2020). A response time process model for not-reached and omitted items. *Journal of Educational Measurement.*

Michaelides, M. P., Ivanova, M., & Nicolaou, C. (2020). The relationship between response-time effort and accuracy in PISA science multiple choice items. *International Journal of Testing, 20*(3), 1-19.

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.

Teig, N., Scherer, R., & Kjærnsli, M. (2020). Identifying patterns of students' performance on simulated inquiry tasks using PISA 2015 log-file data. *Journal of Research in Science Teaching.*

Ulitzsch, E., von Davier, M., & Pohl, S. (2019). A hierarchical latent response model for inferences about examinee engagement in terms of guessing and item-level non-response. *British Journal of Mathematical and Statistical Psychology.*

von Davier, M., Khorramdel, L., He, Q., Shin, H. J., & Chen, H. (2019a). 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., Yamamoto, K., Shin, H. J., Chen, H., Khorramdel, L., Weeks, J., ... & Kandathil, M. (2019b). Evaluating item response theory linking and model fit for data from PISA 2000–2012. *Assessment in Education: Principles, Policy & Practice, 26*(4), 466-488.

Zhang, M., Bennett, R. E., Deane, P., & van Rijn, P. W. (2019). Are there gender differences in how students write their essays? An analysis of writing processes. *Educational Measurement: Issues and Practice, 38*(2), 14-26.