Listening. Learning. Leading.®

Research Report
ETS RR–15-12

Comparing Data Treatments on Item-Level Nonresponse and Their Effects on Data Analysis of Large-Scale Assessments: 2009 PISA Study

Haiwen H. Chen

Matthias von Davier

Kentaro Yamamoto

Nan Kong

June 2015

Discover this journal online at
Wiley Online Library
wileyonlinelibrary.com

ETS Research Report Series

EIGNOR EXECUTIVE EDITOR

James Carlson
Principal Psychometrician

ASSOCIATE EDITORS

Beata Beigman Klebanov
Research Scientist

Heather Buzick
Research Scientist

Brent Bridgeman
Distinguished Presidential Appointee

Keelan Evanini
Managing Research Scientist

Marna Golub-Smith
Principal Psychometrician

Shelby Haberman
Distinguished Presidential Appointee

Donald Powers
Managing Principal Research Scientist

Gautam Puhan
Principal Psychometrician

John Sabatini
Managing Principal Research Scientist

Matthias von Davier
Senior Research Director

Rebecca Zwick
Distinguished Presidential Appointee

PRODUCTION EDITORS

Kim Fryer
Manager, Editing Services

Ayleen Stellhorn
Editor

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Comparing Data Treatments on Item-Level Nonresponse and Their Effects on Data Analysis of Large-Scale Assessments: 2009 PISA Study

Haiwen H. Chen, Matthias von Davier, Kentaro Yamamoto, & Nan Kong

Educational Testing Service, Princeton, NJ

One major issue with large-scale assessments is that the respondents might give no responses to many items, resulting in less accurate estimations of both assessed abilities and item parameters. This report studies how the types of items affect the item-level nonresponse rates and how different methods of treating item-level nonresponses have an effect on item calibration and scoring.

Keywords Programme for International Student Assessment (PISA); item-level nonresponse; low-stakes test; item calibration and scoring; multiple-choice items and constructed items

doi:10.1002/ets2.12059

The Programme for International Student Assessment (PISA) was launched in 2000 by the Organisation for Economic Co-operation and Development to measure how well 15-year-old students are prepared to meet the challenges of today's knowledge societies.

PISA is administered every 3 years. For each assessment, one of the three domains — science, reading, or mathematics — is chosen as the major domain and given greater emphasis on a rotated basis. The remaining two are considered minor domains and are assessed less thoroughly.

Like many large-scale assessments, PISA is a low-stakes test, meaning participants bear little or no consequence for test performance; scores for PISA are not even reported back to the participants. The major issue for low-stakes tests is that some participants may not give their best efforts. They either rapidly make transitions between items, randomly select answers, or skip questions altogether. Results from tests with these types of patterns have been shown to be unreliable (Kiplinger & Linn, 1993; Parshall, 2002).

Many methods have been developed to improve the data quality of large-scale assessments. They fall into two categories: increasing the motivation of the participants and eliminating unreliable responses. Interested readers can get more details in Wise and DeMars (2005) and Lee and Chen (2011).

Reducing the impact of unreliable responses is always a major goal for low-stakes assessments. Researchers have focused on two potential solutions to this problem: identifying and removing unreliable responses and modeling the response rate in latent models to reduce its impact.

Identifying unreliable responses and respondents and removing the data are a direct method of improving data quality. There are several types of unreliable data and methods to remove the data. Because this report will only study item-level nonresponse, the literature on this subject will be omitted. Interested readers can check a detailed survey given by Lee and Chen (2011).

Conversely, some researchers treat the issue of item-level nonresponses as a latent response propensity and incorporate it within a multidimensional item response theory (MIRT) framework. One should note that, unlike the first approach, the MIRT approach does not handle random or quick guessing responses directly, making it less powerful in that regard. Rose, von Davier, and Xu (2010) used several models to examine how MIRT models can reduce the negative impact of item-level nonresponses on simulated and real data from the PISA 2006 test. They included a latent regression model where ability is regressed on the observed response rate of a person; a between-item multidimensional model with two latent variables — one being the ability and the other the response propensity; and a within-item multidimensional IRT model with the same two latent variables. They compared the results from these models with a traditional unidimensional

Corresponding author: H.H. Chen, E-mail: HChen@ets.org

IRT model where item-level nonresponses were either removed or scored as incorrect. Comparisons were made for the simulated data as well as the PISA 2006 data. They found that, with the simulated data, the simple IRT model with item-level nonresponses removed and all three models with the latent response propensity recovered both the item parameters and ability parameters quite well. Conversely, the IRT model with item-level nonresponses recorded as incorrect inflated the item difficulties dramatically.

There is an additional issue for PISA. The calibration and scoring procedures for PISA up to 2012 were done in two steps: First, the item parameters were estimated using an IRT model with certain item-level nonresponses removed; second, with all item-level nonresponses treated as incorrect, abilities of the students were estimated using the same IRT model as where the item parameters from Step 1 were fixed. As Rose et al. (2010) pointed out, the PISA procedures change the data set between-item calibration using IRT (where item-level nonresponses were removed) and ability estimation using latent regression-based population models (where item-level nonresponses were scored as incorrect), thus changing basic item statistics such as the percentage correct and item total correlations between different stages of the analysis. They demonstrated that using simulated data with an item-level nonresponse rate of 30% can result in a mean ability 0.6 logits lower than if the item-level nonresponses were removed for both calibration and scoring.

The simulation results suggested that removing all item-level nonresponses for both item calibration and scoring can improve the accuracy for both item parameters and ability variables, but this may not be the case for real data. The most acceptable approach is likely somewhere in between — recoding some of the item-level nonresponses as incorrect while removing others (based on some criterion).

With the administration of PISA 2015 approaching, one focus is on reducing unreliable responses and, as a result, increasing the quality of the assessment. According to Rubin's framework (Little & Rubin, 2002; Rubin, 1976), some data are missing at random (MAR), and removing them will improve the quality of the data analysis; others are missing not at random (MNAR), and removing them directly may even impair the data quality. The goal of this study is to determine how likely any missing data are MAR by examining the impact of item-level nonresponses and to find methods to reduce the impact, thus improving the design and implementation of PISA 2015.

Using data from PISA 2009, this report studies two issues of item-level nonresponses — first, how the types of items affect the item-level nonresponse rates and, second, how different methods of treating item-level nonresponses have an effect on item calibration and scoring. The result may give us some idea how to improve the design and implementation of PISA 2015.

The rest of the paper is organized as follows: First we analyze the distributions of item-level nonresponses among all item types. Then, several item calibration and scoring methods with different treatments on item-level nonresponses are compared to see which method will produce more reasonable results for large-scale assessment. We end with a concluding discussion.

Item-Level Nonresponses Versus Item Types

Most international assessments distinguish between two types of item-level nonresponse: missing and not reached. Missing responses are those that are followed by at least one observed response. Not reached are item-level nonresponses grouped at the end of the assessment or a section of the assessment that are not followed by any observed responses within the section. It is typically assumed for not reached that the test taker stopped responding because of time limits or just quit the test. It appears important to note that the amount of missingness and the propensity to respond vary across test takers as well as across countries. Potential reasons for this variation on the individual level are, among others, achievement motivation (Eccles, Wigfield, & Schiefele, 1998; Wise & DeMars, 2005) and country-level variables, such as familiarity with standardized testing situations (e.g., Cosgrove, 2011; Nunnally, 1967).

The major domain in PISA 2009 was reading. The assessment consisted of 131 reading items, 53 science items, and 35 math items, which were used to generate 20 test booklets of different difficulty levels. The average number of items in the booklets was 60 (Programme for International Student Assessment, 2009). Each student was randomly assigned one booklet.[1] In 2009, there were 470,000 students representing 65 nations. An additional 50,000 students representing nine nations and regions were tested in 2010 using the same set of questions (Organisation for Economic Co-operation and Development, 2012).

Item types used in PISA 2009 were either selected response or constructed response. Selected-response items were either standard multiple-choice items, for which students were required to select the correct answer, or complex

multiple-choice items, for which each item had several true–false types of questions grouped together (yes–no, true–false, correct–incorrect, etc.).

Constructed-response items were of three broad types. Closed-constructed-response items required students to construct a numeric response within very limited constraints or only required a word or short phrase as the answer. Short-response items required a response generated by the student, with a limited range of possible full-credit answers. Open-constructed-response items required more extensive writing and frequently required some explanation or justification. In PISA 2009, for 219 math, reading, and science items, the breakdown was as follows: 79 multiple-choice items, 34 complex multiple-choice items, 17 closed-constructed items, 19 short-response items, and 70 open-constructed items.

The rate of one particular response (correct, incorrect, missing, or not reached) on an item type for a country or region (or all countries and regions) is the ratio of the number of all such responses versus the number of all items (one item counted N times if it was given to N students) within the specified item type given to the specified group of students. In this way, all responses have equal weight. The rate should be equivalent to the average rate of all items in the same item group if all items in the group are evenly given to participants.

The question of interest is how item types affect both the missing and not reached rates. Table A1 shows the not reached rates for all countries and regions grouped by item types. In addition to the five item types previously noted, the overall not reached rate is listed, as is the max–min rate, which is the difference between the maximum and minimum not reached rates of all five item types in a country or region. For example, for Argentina, the maximum not reached rate is 10.4% (closed constructed), and the minimum is 9.4% (short response), with a resulting max–min of 1.0%.

The minimum not reached rates are as low as 0.1% in some countries and regions, and the maximum rates are as high as 16% in others. Yet the difference between the maximum and minimum for a particular country or region is never more than 2%. This should not be surprising because any item in PISA 2009 was assigned to five to eight booklets, and its positions in all booklets were evenly distributed. In particular, there is equal probability that any item type will appear as not reached.

Table A2 provides the same types of information for the missing rates. The ranges of missing rates can be as low as 0.4% in one country or region and as high as 45.9% in another. Note that the missing rates have a far different pattern. The majority of countries and regions have max–min rates more than twice the average of their missing rates. In other words, in contrast to the not reached rates, missing rates strongly depend on the item types. The missing rates, as one might expect, are higher for constructed-response items, with open-constructed items having the highest rates, followed by short-response and closed-constructed items. Complex multiple choice is the lowest, followed by multiple choice.

To make the item-type analysis complete, we also include the tables for correct and incorrect rates, where item-level nonresponses were excluded. The counting of correct or incorrect responses is straightforward for dichotomous items. For polytomous items, if the examinee received full credit, the response is treated as one correct response; if the examinee received zero credit, the response is treated as one incorrect response; if the examinee received partial credit, then the response is treated as a partial correct response and a partial incorrect response, determined by the ratio of credit received. For example, if the full credit is 2 for a polytomous item and a student's score is 1, then we have one-half of a correct response and one-half of an incorrect response. In this way, the counting is consistent among all response types (missing, not reached, correct, and incorrect). For PISA 2009, there were 14 polytomous items, and all of them have full credit of 2.

As one can see in Table A3, the ranges of the correct rates vary significantly among countries and regions. In general, the maximum rate is about 3 times the minimum on the same item type among all countries and regions. There are also quite large differences in average correct rates among the item types. The average percentage correct for multiple-choice and closed-constructed items for all countries and regions is about 15 percentage points higher than for the other three types of items. The average percentage correct for complex multiple-choice items for all countries and regions is the lowest, although the average missing rate for complex multiple-choice items is also the lowest in the group of item types.

Also, as can be seen in Table A4, the variation of the incorrect rates is about the same as for the missing rates and correct rates. The average percentage incorrect for all item types except for complex multiple-choice items is in the range of 30%–40%, whereas the rate for complex multiple-choice items is 53.5%, 13.7% more than the item type with the next highest percentage incorrect.

Data Analysis With Options on Item-Level Nonresponses

The main focus of this study is to compare the impact of different approaches to coding item-level nonresponses on the estimation of both item parameters and student ability. Because there are two kinds of item-level nonresponses, we conduct

Table 1 Statistics of the Mean Weighted Likelihood Estimates Abilities Under Three Conditions on Item-Level Nonresponses

Country/region	Condition A	Condition B	Condition C	Cond. B – Cond. A	Cond. C – Cond. A
Mean	0.204	0.293	0.490	0.089	0.286
SD	0.643	0.625	0.570	0.051	0.124

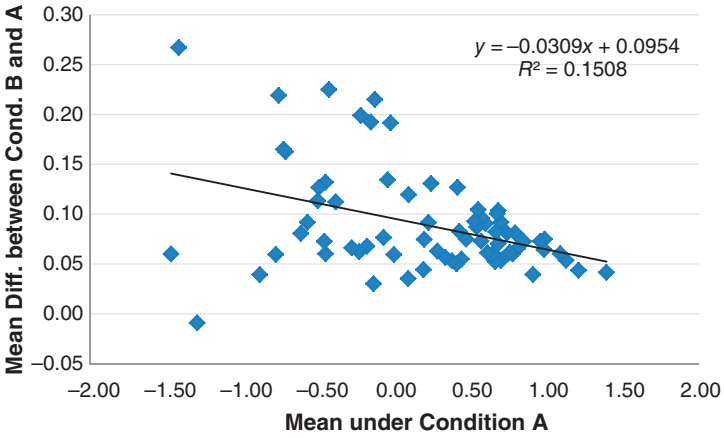

Figure 1 Plot of mean difference between Conditions B and A versus mean under Condition.

the analysis under the following three conditions: treat all item-level nonresponses as incorrect (Condition A); treat missing as incorrect and remove not reached responses (Condition B); and remove all item-level nonresponses (Condition C).

The software ConQuest was used with the Rasch model for item calibration. This is the same program that was used for PISA 2009 operationally. ConQuest uses marginal maximum likelihood methods to obtain item parameter and population distribution estimates for the mixed-coefficients multinomial logit model (Adams & Wu, 2007), which is a generalized form of the Rasch model.

The ConQuest software offers the option to produce weighted likelihood estimates (WLE; Warm, 1989), which were used in this study. Compared to maximum likelihood estimates, which is also used in ConQuest, WLE offers a bias reduction and also provides meaningful estimates for response patterns with extreme scores (all zero or all correct). The results of the current analysis will not be compared with PISA 2009 results, where item calibration was performed with all not reached item-level nonresponses removed but scoring was done with all item-level nonresponses treated as incorrect. Such a comparison should produce results similar to those presented in Rose et al. (2010). Because of the large data sets, the calibrations and scorings were performed on each domain (math, reading, and science) separately to save computing time.

Because reading was the major domain for PISA 2009, we present our results mainly for that domain.

Of the 515,958 students taking PISA 2009, 514,478 students answered at least one reading item. The number of the reading items can be as low as 11 in one booklet and as high as 61 in another booklet. The not reached rate of reading items across all countries is 3.3%, and the missing rate is 8.9%. These are similar to the rates for items in all three domains (3.2% not reached and 9.3% missing; see Tables A1 and A2). The item calibrations and scorings were conducted simultaneously under each condition. To compare the ability changes under different conditions, we set the sum of all item difficulties as zero for all three conditions.

Table A5 lists the mean WLE abilities for all countries and regions under the three conditions on item-level nonresponses, sorted by the means under Condition A. The differences in means under Condition A versus Conditions B and C are also included.

The (unweighted) means and standard deviations of the country or region means and the differences under the three conditions are given in Table 1.

The means for all countries and regions under Condition A and the mean differences between Conditions B and A are plotted in Figure 1. The same means and the mean differences between Conditions C and A are plotted in Figure 2.

The negative correlations between the ability and the increases in ability estimates by removing certain item-level nonresponses can be explained as follows: The lower the average ability is of the participating students in a country or region,

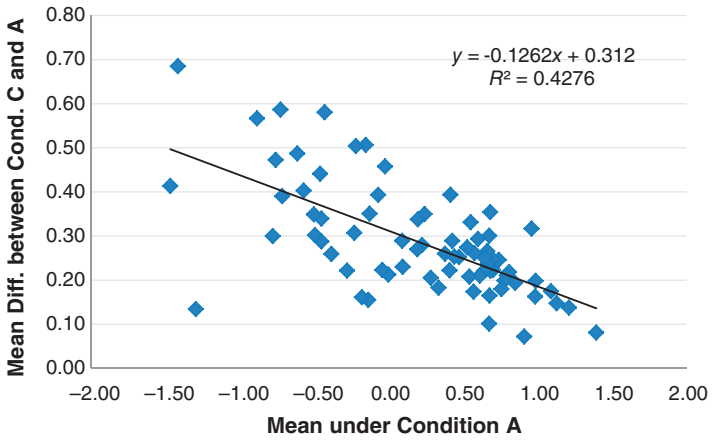

Figure 2 Plot of mean difference between Conditions C and A versus mean under Condition.

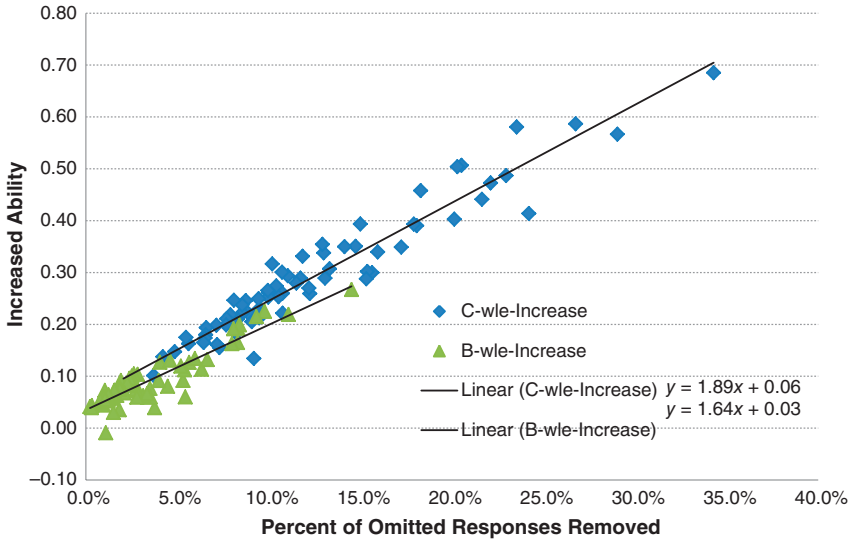

Figure 3 Relationships between increased ability and percentage of certain item-level nonresponses removed.

the better the chance is of a higher percentage of item-level nonresponses. The correlation became stronger when all item-level nonresponses were removed. This also explains that the standard deviations get smaller from Condition A to B and on to C.

Figure 3 shows that the gains are highly correlated with the item-level nonresponse percentages. Because mean ability increases were not consistent across all countries and regions after recoding certain item-level nonresponses, it can be difficult to discern the relative impact of the three approaches. As such, changes in country ranks are more informative. Figures A1 and A2 plot the rank gain (loss) under Conditions B and C from Condition A, respectively.

The largest rank change under Condition B is −4 for Mauritius, but for Condition C, it is −10 for the United States.

To see how removing certain item-level nonresponses affects the item parameter estimation, we plot the difficulty parameter values under Condition A versus Condition B in Figure 4 and Condition A versus Condition C in Figure 5.

Because the sum of parameter values for all items is set to zero in the calibration, item parameter estimates under any two different conditions are quite similar. However, there are some moderate deviations in the estimates between Conditions A and C.

Discussion

Item-level nonresponses resulting from a lack of test-taking motivation are always a major issue for large-scale assessments. Such responses reduce the accuracy of item calibrations and scoring. Reducing these types of responses can be

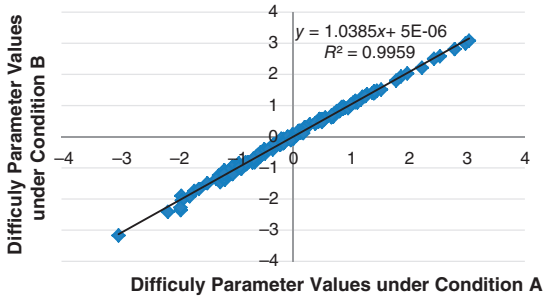

Figure 4 Correlation between item parameters under Condition A versus Condition B.

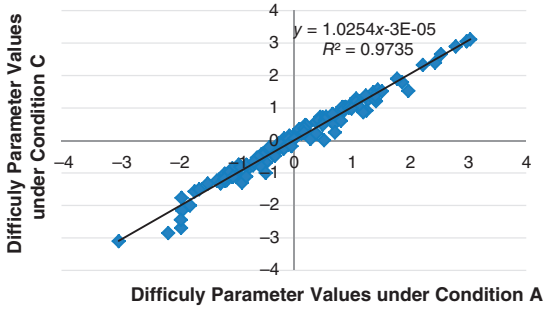

Figure 5 Correlation between item parameters under Condition A versus Condition C.

accomplished by improving the test design, eliminating associated responses, and/or motivating the respondents. The current study mainly addresses the second approach but also raises some issues about the test design.

Table A1 shows that the assessment design is quite balanced on item types, particularly at the end of all sections of the booklets. Conversely, Table A2 indicates that students more likely skipped the constructed-response than the multiple-choice questions, although some multiple-choice questions (such as complex multiple-choice questions) are actually harder than some constructed-response questions (such as closed-constructed questions), as Table A3 shows. It is not surprising to see complex multiple choice with a much higher incorrect rate than all other item types.

This study reconfirms research results that constructed-response items will get a higher item-level nonresponse rate than multiple-choice items (see DeMars, 2000). In PISA 2009, about 50% of items were constructed response. In particular, one third of constructed-response items were open constructed. Because the not reached rates for PISA 2009 were quite alike for all item types, our focus is on the missing rates.

There are two possible ways to reduce the impact of the missing rates. One is to reduce the percentage of constructed-response items, under the condition that the coverage on the contents is satisfied. Another is to use more closed-constructed items, because the missing rate of closed-constructed items is half that of the other two constructed-response (short response and open constructed) items. Considering that the percentage of closed-constructed items was less than 8% in PISA 2009, the second approach may be more feasible.

The overall correct rate for PISA 2009 was below 48%, which is quite low for a large-scale assessment. One would investigate the items to see why some items had very low correct rates, particularly the complex multiple-choice items, which had both the lowest missing rate and the lowest correct rate. Because the question (true–false) type is quite easy for the complex multiple-choice item type, there is a tendency to increase the item difficulty by using a long reading format, placing the question in a distant position, and/or increasing the number of questions for each item. Test designers need to reexamine the items to make the correct rates of all item types suitable for the assessment.

Removing unreliable data may improve the assessment, but how to determine unreliable data is a tricky issue. Because PISA was primarily a paper-and-pencil assessment, the only unreliable data we can possibly identify are within item-level nonresponses. From the item analysis on item-level nonresponses, one can see that using Rubin's framework (Little & Rubin, 2002; Rubin, 1976), missing is definitely MNAR, because more difficult item types in perception have higher missing rates, whereas not reached is more likely MAR, based on the assumptions that the item types were randomly

assigned at the end of booklets (confirmed) and the participants may not have seen the questions, either intentionally or because they did not have time.

The comparison of data analyses shows that by making the sums of item difficulties as 0 under both Condition A (treat both not reached and missing as incorrect) and Condition B (remove not reached and treat missing as incorrect), the item difficulties are highly correlated (Figure 4). With the assumption that the test design has balanced the item difficulties with item locations, shortening the test for randomly chosen students (Condition B) will not change the item parameters and student abilities theoretically. Therefore, the result under Condition B may be closer to the result obtained if there were no missing data than under Condition A. Condition B benefits countries and regions with larger not reached rates. However, most countries and regions have not reached rates of less than 10%. The biggest change in ranking from Condition A to Condition B is −4 for Mauritius, although its ability increase under Condition B is the same as the average of all countries and regions.

Condition C (remove both not reached and missing) made the assumption that all item-level nonresponses are MAR, which is definitely not true. The item-level nonresponse rates on item types clearly show that the constructed-type items had much higher item-level nonresponse rates. The item parameters estimated under both Condition A and Condition C show they are not comparable, surprisingly on easy items (Figure 5). Although we do not know the answer, it is likely that an increased level of missing is associated with a lower expected ability, both on the individual and on the country level. More specifically, students with lower ability tend to skip more questions in the assessment. However, the association is neither perfect nor deterministic (Rose et al., 2010), and it can be moderated by other variables such as achievement motivation (Eccles et al., 1998; Wise & DeMars, 2005). Interestingly, some high-performance countries get the most benefit from the treatment of the item-level nonresponses on Condition C. The country that jumps the highest in rank is France. Its improvement is 0.354 logits with respect to mean ability, and it moves from 18th place under Condition A to 9th place under Condition C. Japan is another high-performance country, with an improvement in mean ability of 0.317 logits and a change in rank from seventh place to third.

The findings in this report suggest if we do item calibration and scoring under Condition B, the item difficulty parameters and the scoring may be more accurate than under either Condition A or Condition C. However, we need to check and revise the test instruction to ensure that it does not encourage intentionally skipping the later questions.

The item analysis indicates that the countries and regions with students of lower ability tend to skip more questions than the countries and regions where students are more able, and the most skipped item types are open constructed and short response. Because the design of PISA intends to give lower performing countries and regions easier questions, perhaps it is possible to give them more items that typically have fewer missing types, such as multiple choice.

This study is an initial step in exploring the options if a different data calibration and scoring method can be used. Further study with the finer treatments on nonresponses may give better estimations of the item parameters and country and region performances. Some possible scenarios are removing a certain percentage of missing responses and/or removing missing responses of certain types (suggested by a reviewer). Of course, if PISA completely transitions to computer-based assessment, we will have much more efficient means to reduce and even eliminate the impact of unreliable data.

Acknowledgments

The authors thank Lale Khorramdel, Jonathan Weeks, Daniel McCaffrey, Yue Jia, and Shelby Haberman for their exceptionally valuable comments and suggestions. The authors' appreciation also goes to Larry Hanover for his excellent editing of this report.

Note

1 Certain (easier) booklets were given to nations and regions that had lower performance. These nations and regions were predetermined by PISA 2006 or the 2009 PISA Field Trial.

References

Adams, R. J., & Wu, M. L. (2007). The mixed-coefficients multinomial logit model: A generalized form of the Rasch model. In M. von Davier, & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models—Extensions and applications* (pp. 57–75). New York, NY: Springer.

Cosgrove, J. (2011). *Does student engagement explain performance on PISA? Comparisons of response patterns on the PISA tests across time*. Dublin, Ireland: Educational Research Centre.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education*, *13*(1), 55 – 77.

Eccles, J. S., Wigfield, A., & Schiefele, U. (1998). Motivation to succeed. In W. Damon (Series Ed.) and N. Eisenberg (Vol. Ed.), *Handbook of child psychology* (Vol. 3, 5th ed., pp. 1017 – 1095). New York, NY: Wiley.

Kiplinger, V. L., & Linn, R. L. (1993). *Raising stakes of test administration: The impact on student performance on NAEP* (Technical report). Los Angeles, CA: Center for Research on Evaluation, Standards, and Student Testing.

Lee, Y., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*(3), 359 – 379.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Hoboken, NJ: Wiley-Interscience.

Nunnally, J. (1967). *Psychometric theory*. New York, NY: McGraw-Hill.

Organisation for Economic Co-operation and Development. (2012). *PISA 2009 technical report*. Retrieved from http://www.oecd.org/pisa/pisaproducts/pisa2009/50036771.pdf

Parshall, C. (2002). Item development and pretesting in a CBT environment. In C. Mills, M. Potenza, J. Fremer, & W. Ward (Eds.), *Computer-based testing: Building the foundation for future assessments* (pp. 119 – 141). Mahwah, NJ: Lawrence Erlbaum.

Programme for International Student Assessment. (2009). *PISA 2009 codebook*. Retrieved from http://pisa2009.acer.edu.au/downloads/Codebook_COG09_TD_DEC11.pdf

Rose, N., von Davier, M., & Xu, X. (2010). *Modeling nonignorable missing data with item response theory (IRT)* (Research Report No. RR-10-11). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02218.x

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581 – 592.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427 – 450.

Wise, S., & DeMars, C. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1 – 17.

Appendix

Tables and Graphs of Item Type and Country/Region Information

Table A1 Not Reached Rates of Item Types by Country or Region

	Rate (%)						
Country/region	All	MC	CMC	Closed	Short	Open	Max – min
Albania	4.7	4.6	4.6	5.2	4.6	4.6	0.6
UAE	2.7	2.8	2.8	2.5	2.4	2.6	0.4
Argentina	9.9	10.2	10.0	10.4	9.4	9.6	1.0
Australia	1.9	2.0	1.8	2.1	1.9	1.8	0.3
Austria	1.0	1.1	0.8	1.3	0.9	0.9	0.5
Azerbaijan	2.3	2.5	1.3	2.5	2.5	2.6	1.4
Belgium	1.6	1.8	1.6	1.7	1.3	1.4	0.5
Bulgaria	3.7	3.5	3.7	4.4	3.5	3.6	0.8
Brazil	5.3	5.3	5.2	5.7	5.3	5.1	0.5
Canada	1.5	1.6	1.4	1.7	1.4	1.3	0.3
Switzerland	1.1	1.2	1.1	1.3	1.1	1.0	0.3
Chile	3.7	3.9	3.7	3.6	3.5	3.5	0.4
Colombia	9.3	9.7	9.6	9.0	9.0	9.0	0.8
Costa Rica	5.1	5.5	5.4	4.4	5.0	4.9	1.1
Czech Republic	1.0	1.2	1.0	1.3	1.1	0.9	0.4
Germany	1.2	1.4	1.1	1.4	1.1	1.1	0.3
Denmark	1.8	1.9	1.7	2.0	1.8	1.6	0.4
Spain	2.5	2.6	2.4	2.9	2.5	2.3	0.6
Estonia	1.1	1.2	0.9	1.2	1.0	0.9	0.3
Finland	0.8	1.0	0.7	1.0	0.9	0.7	0.3
France	3.0	3.3	2.7	3.2	3.0	2.7	0.6
United Kingdom	1.1	1.2	1.0	1.2	1.0	1.0	0.3
Georgia	7.9	8.2	7.2	8.8	8.0	7.9	1.7
Greece	2.7	2.8	2.7	3.4	2.4	2.5	1.0

Table A1 Continued

	Rate (%)						
Country/region	All	MC	CMC	Closed	Short	Open	Max – min
Hong Kong — China	0.7	0.7	0.5	1.0	0.7	0.6	0.5
Croatia	0.8	0.8	0.6	1.1	0.7	0.7	0.5
Hungary	0.8	0.8	0.7	1.1	0.8	0.7	0.4
Indonesia	5.2	5.3	5.2	6.0	4.9	5.0	1.1
Ireland	1.8	1.9	1.6	2.3	1.8	1.7	0.7
Iceland	2.1	2.2	2.0	2.3	2.1	1.9	0.4
Israel	4.0	4.3	3.7	4.6	3.7	3.7	1.0
Italy	2.2	2.4	2.1	2.6	2.1	2.0	0.5
Jordan	3.4	3.5	3.6	3.4	3.2	3.3	0.4
Japan	1.2	1.2	1.1	1.5	1.3	1.1	0.3
Kazakhstan	6.3	6.4	6.2	6.4	6.3	6.1	0.4
Kyrgyzstan	15.1	15.1	15.1	16.3	14.9	14.9	1.4
Korea	0.3	0.3	0.2	0.4	0.3	0.3	0.1
Liechtenstein	1.4	1.4	1.3	2.0	1.3	1.4	0.8
Lithuania	1.1	1.2	1.0	1.1	1.0	0.9	0.3
Luxembourg	2.3	2.5	2.1	2.7	2.1	2.1	0.7
Latvia	1.5	1.7	1.4	1.7	1.5	1.3	0.3
Macao — China	2.1	2.3	2.1	2.3	1.9	1.8	0.5
Moldova	4.9	5.1	4.7	5.3	4.6	4.9	0.7
Mexico	5.9	6.2	6.0	5.4	5.5	5.6	0.8
Malta	2.7	2.8	2.5	3.0	2.4	2.8	0.6
Montenegro	2.6	2.9	2.4	3.0	2.6	2.4	0.7
Mauritius	3.6	3.7	3.6	3.6	3.4	3.5	0.3
Malaysia	2.8	3.0	2.8	2.8	2.6	2.8	0.4
Netherlands	0.3	0.3	0.3	0.3	0.3	0.3	0.1
Norway	1.9	2.1	1.8	2.2	1.8	1.7	0.4
New Zealand	1.7	1.9	1.6	1.9	1.7	1.6	0.3
Panama	7.8	8.2	7.5	7.7	7.8	7.5	0.7
Peru	11.1	11.3	10.9	11.7	11.0	10.8	1.0
Poland	0.9	1.0	0.8	1.5	1.1	0.9	0.7
Portugal	2.4	2.6	2.4	2.8	2.4	2.2	0.6
Qatar	3.4	3.3	3.2	4.0	3.2	3.5	0.8
Shanghai — China	0.2	0.2	0.1	0.3	0.2	0.2	0.2
Himachal Pradesh — India	4.7	4.7	4.0	6.1	4.6	4.9	2.0
Tamil Nadu — India	1.0	1.0	0.8	1.5	1.1	1.0	0.6
Miranda — Venezuela	7.9	8.2	8.0	7.4	7.6	7.6	0.8
Romania	1.3	1.4	1.3	1.4	1.2	1.3	0.2
Russia	4.4	4.6	4.3	5.0	4.4	4.1	0.8
Singapore	1.2	1.3	1.2	1.2	1.2	1.1	0.1
Serbia	1.9	2.0	1.9	1.6	1.7	1.8	0.3
Slovak	1.1	1.1	0.9	1.4	1.1	1.0	0.4
Slovenia	0.8	1.0	0.6	1.1	0.8	0.7	0.5
Sweden	2.6	2.7	2.4	3.0	2.4	2.4	0.6
Chinese Taipei	0.8	0.8	0.7	1.2	0.8	0.8	0.5
Thailand	2.0	2.1	2.1	2.4	2.0	1.9	0.5
Trinidad and Tobago	8.2	8.6	8.1	7.9	7.6	8.1	1.0
Tunisia	6.1	6.3	6.0	6.3	5.7	6.1	0.6
Turkey	1.5	1.7	1.5	1.8	1.6	1.3	0.5
Uruguay	8.0	8.4	7.9	8.0	7.8	7.8	0.6
United States	1.0	1.0	0.9	1.2	0.9	0.9	0.2
Average	3.2	3.3	3.1	3.4	3.1	3.1	0.4
Min	0.2	0.2	0.1	0.3	0.2	0.2	0.1
Max	15.1	15.1	15.1	16.3	14.9	14.9	2.0

Note. Here max – min is the difference between the maximum not reached rate minus the minimum not reached rate for all item types for a given country or region. MC = multiple choice; CMC = complex multiple choice; Closed = closed-constructed response; Short = short response; Open = open-constructed response.

Table A2 Missing Rates of Item Types by Country or Region

	Rate (%)						
Country/region	All	MC	CMC	Closed	Short	Open	Max – min
Albania	20.8	9.5	10.5	20.5	29.0	35.1	25.6
UAE	7.1	3.3	2.4	7.1	10.4	12.5	10.1
Argentina	16.0	8.2	7.3	12.5	25.1	26.7	19.4
Australia	5.9	1.9	1.7	4.8	10.1	11.7	10.0
Austria	9.0	3.4	2.7	6.7	11.2	18.6	15.9
Azerbaijan	23.0	5.6	7.3	25.0	31.9	45.9	40.3
Belgium	6.4	2.6	1.7	4.0	8.5	13.0	11.3
Bulgaria	15.0	4.6	6.1	13.2	22.3	28.6	24.0
Brazil	8.0	2.8	4.0	7.0	14.0	13.6	11.2
Canada	4.8	1.5	1.3	3.6	8.1	9.7	8.4
Switzerland	6.8	2.4	1.7	4.6	8.1	14.5	12.8
Chile	9.1	2.8	2.9	7.6	17.7	16.7	14.9
Colombia	7.3	2.6	3.2	6.5	13.4	12.6	10.8
Costa Rica	5.5	2.3	2.0	4.6	9.0	9.8	7.7
Czech Republic	8.9	2.9	2.1	6.2	13.1	18.8	16.7
Germany	7.9	2.7	2.0	5.3	10.3	16.7	14.7
Denmark	8.6	2.7	2.6	6.6	12.0	17.8	15.2
Spain	8.4	2.9	2.4	7.2	13.8	16.6	14.2
Estonia	5.5	1.6	1.0	3.2	7.6	12.1	11.1
Finland	4.4	1.3	1.1	2.9	6.3	9.5	8.4
France	9.7	3.7	2.4	5.9	12.6	20.4	18.0
United Kingdom	7.0	2.5	1.8	4.8	11.0	14.0	12.2
Georgia	20.9	8.4	11.0	20.5	31.0	36.0	27.7
Greece	10.2	3.3	3.3	8.9	16.3	20.3	17.1
Hong Kong — China	4.0	1.0	0.7	2.4	5.3	9.1	8.4
Croatia	9.0	2.9	2.2	6.2	13.9	18.5	16.2
Hungary	7.7	1.5	1.3	5.0	10.9	17.4	16.0
Indonesia	11.0	4.6	5.7	10.3	19.8	18.9	15.2
Ireland	6.6	2.5	1.6	5.3	10.3	13.1	11.4
Iceland	6.2	2.4	1.8	4.4	9.2	12.5	10.7
Israel	11.5	5.2	4.2	11.0	16.6	21.6	17.5
Italy	9.0	2.7	2.4	6.7	12.5	19.1	16.8
Jordan	8.1	3.6	3.4	7.6	11.5	14.2	10.8
Japan	8.3	1.7	1.2	6.0	12.0	18.8	17.7
Kazakhstan	12.1	4.6	5.5	11.2	17.8	21.9	17.3
Kyrgyzstan	20.4	8.1	11.4	20.8	30.2	34.7	26.6
Korea	3.7	0.8	0.6	2.2	4.4	8.6	8.0
Liechtenstein	6.4	1.7	1.4	3.5	7.8	14.5	13.2
Lithuania	7.1	1.8	1.5	4.8	11.1	15.4	13.9
Luxembourg	9.2	3.7	2.7	6.8	12.0	18.6	15.9
Latvia	5.4	1.5	1.0	3.6	9.2	11.4	10.4
Macao — China	5.3	1.5	1.1	2.7	6.4	11.8	10.8
Moldova	16.6	6.7	7.2	15.2	24.7	29.7	23.0
Mexico	4.7	1.9	1.9	4.0	7.6	8.3	6.4
Malta	10.3	4.1	2.7	8.6	16.0	19.5	16.8
Montenegro	19.4	7.5	6.6	17.5	28.8	36.8	30.2
Mauritius	11.7	3.9	4.1	11.2	18.9	21.7	17.9
Malaysia	11.0	3.0	3.9	12.1	17.1	20.9	17.9
Netherlands	1.9	0.6	0.5	1.2	3.6	4.0	3.5
Norway	7.5	2.7	2.3	5.7	11.7	14.9	12.6
New Zealand	5.2	1.9	1.4	3.8	8.4	10.6	9.2
Panama	11.4	6.3	4.9	11.1	17.2	18.5	13.7
Peru	13.2	7.3	7.3	12.9	20.8	20.5	13.5
Poland	6.8	1.3	1.1	4.2	9.0	15.7	14.6
Portugal	7.7	1.9	1.5	5.0	11.9	16.9	15.4
Qatar	12.4	4.2	4.8	13.5	19.2	22.5	18.3
Shanghai — China	2.0	0.4	0.4	1.3	3.2	4.5	4.1
Himachal Pradesh — India	20.5	11.1	10.2	21.8	25.7	34.8	24.5

Table A2 Continued

	Rate (%)						
Country/region	All	MC	CMC	Closed	Short	Open	Max – min
Tamil Nadu — India	9.0	4.6	5.2	9.3	14.1	14.4	9.8
Miranda — Venezuela	12.3	6.9	6.0	10.2	18.6	19.7	13.7
Romania	6.1	2.4	2.4	5.4	9.7	11.0	8.6
Russia	9.9	3.4	3.4	7.7	14.0	20.0	16.6
Singapore	4.2	1.6	1.0	2.6	6.1	8.7	7.7
Serbia	12.7	4.8	4.3	9.2	18.7	23.9	19.5
Slovak	9.2	2.2	2.2	7.0	13.3	20.0	17.8
Slovenia	10.9	3.1	2.1	8.3	16.9	23.2	21.1
Sweden	8.1	3.0	3.0	6.6	11.1	15.9	12.9
Chinese Taipei	5.3	1.1	0.9	3.4	8.3	12.1	11.2
Thailand	6.0	3.0	1.8	5.6	10.5	10.5	8.7
Trinidad and Tobago	12.7	6.4	5.5	11.8	18.6	21.6	16.1
Tunisia	10.9	6.1	3.8	10.5	16.1	17.9	14.1
Turkey	7.5	2.2	2.2	6.6	11.9	15.3	13.0
Uruguay	13.4	6.3	5.4	11.6	20.1	23.2	17.8
United States	2.6	0.6	0.9	2.6	5.3	5.1	4.7
Average	9.3	3.5	3.3	7.9	13.9	17.7	14.4
Min	1.9	0.4	0.4	1.2	3.2	4.0	3.5
Max	23.0	11.1	11.4	25.0	31.9	45.9	40.3

Note. MC = multiple choice; CMC = complex multiple choice; Closed = closed-constructed response; Short = short response; Open = open-constructed response.

Table A3 Correct Rates of Item Types by Country or Region

	Rate (%)						
Country/region	All	MC	CMC	Closed	Short	Open	Max – min
Albania	35.2	46.0	26.0	44.8	31.7	28.1	20.0
UAE	44.7	55.5	34.9	55.8	40.1	38.0	20.9
Argentina	37.7	47.4	28.4	49.6	33.2	31.4	21.1
Australia	57.2	64.1	52.0	65.3	49.4	52.7	15.9
Austria	52.9	59.8	50.3	62.5	50.1	45.1	17.4
Azerbaijan	32.9	45.9	23.3	45.9	36.7	21.1	24.9
Belgium	58.1	64.9	51.1	69.0	53.6	53.2	17.9
Bulgaria	44.3	55.3	33.6	53.5	40.9	37.7	21.8
Brazil	36.6	48.2	27.3	45.3	32.0	29.6	20.9
Canada	57.9	64.8	50.7	66.1	52.3	54.0	15.4
Switzerland	56.5	62.8	51.2	64.6	55.2	50.9	13.8
Chile	47.1	59.5	35.6	56.8	41.5	40.0	23.9
Colombia	39.8	51.3	27.5	50.5	34.5	33.7	23.8
Costa Rica	43.5	55.5	31.0	54.3	40.9	36.5	24.5
Czech Republic	57.1	65.9	51.1	67.6	51.9	49.8	17.7
Germany	57.2	63.9	54.0	65.2	53.3	50.6	14.6
Denmark	51.7	60.4	46.2	61.4	45.7	44.4	17.0
Spain	52.7	60.9	46.4	60.1	47.7	46.7	14.5
Estonia	57.8	63.3	50.8	68.6	54.3	54.1	17.8
Finland	63.3	70.6	57.5	72.1	56.6	57.9	15.5
France	54.8	62.1	49.2	64.7	49.9	48.9	15.8
United Kingdom	54.3	60.6	50.8	63.2	46.3	49.2	16.9
Georgia	32.7	44.2	23.7	40.4	29.8	25.0	20.4
Greece	50.2	58.9	41.6	58.0	40.4	46.2	18.5
Hong Kong — China	64.0	69.1	55.2	75.0	58.2	62.5	19.8
Croatia	49.9	57.5	43.7	62.4	42.3	44.4	20.0
Hungary	55.1	64.7	48.7	62.6	50.6	48.4	16.3
Indonesia	31.8	39.6	23.2	40.6	25.0	27.5	17.4
Ireland	55.1	60.7	50.5	61.9	48.5	51.5	13.4

Table A3 Continued

	Rate (%)						
Country/region	All	MC	CMC	Closed	Short	Open	Max – min
Iceland	55.1	63.6	46.9	65.2	51.6	49.0	18.4
Israel	48.0	56.3	39.7	53.8	39.9	43.8	16.6
Italy	53.5	62.8	47.0	61.6	48.2	46.3	16.5
Jordan	39.8	51.6	28.3	46.8	34.4	34.1	23.3
Japan	61.0	66.6	56.0	71.6	53.9	56.9	17.6
Kazakhstan	37.0	49.5	27.7	47.1	35.2	27.8	21.7
Kyrgyzstan	24.0	34.3	17.0	29.4	20.9	17.1	17.3
Korea	64.1	72.5	54.0	75.7	57.8	59.4	21.7
Liechtenstein	57.8	64.3	52.3	63.9	56.2	52.8	12.0
Lithuania	50.6	57.7	42.3	63.3	46.4	45.6	20.9
Luxembourg	51.8	58.7	47.4	60.7	48.1	45.6	15.1
Latvia	53.2	60.3	45.4	61.0	46.2	49.7	15.6
Macao — China	55.1	61.6	47.9	66.4	52.2	50.3	18.5
Moldova	36.7	44.8	28.5	46.6	34.1	31.6	18.1
Mexico	42.6	54.1	31.6	55.7	42.0	34.2	24.1
Malta	50.3	59.9	42.1	61.4	46.3	43.6	19.3
Montenegro	35.3	45.2	28.5	48.9	27.9	27.2	21.7
Mauritius	38.3	49.7	29.8	47.8	35.2	30.0	19.9
Malaysia	40.9	54.8	29.9	49.6	34.4	32.6	24.9
Netherlands	60.4	67.3	54.5	69.2	56.7	55.1	14.7
Norway	55.4	63.0	49.8	62.3	51.0	49.7	13.3
New Zealand	60.1	66.3	55.4	68.1	52.7	56.0	15.3
Panama	32.2	42.3	24.0	38.8	28.8	26.0	18.2
Peru	31.9	42.9	23.4	41.8	28.2	24.1	19.5
Poland	56.8	65.0	50.7	67.4	50.3	50.4	17.1
Portugal	52.6	60.5	45.8	61.3	46.8	47.1	15.5
Qatar	32.5	44.0	25.8	40.5	27.0	24.6	19.4
Shanghai — China	69.3	74.9	56.8	79.4	65.8	68.7	22.6
Himachal Pradesh — India	23.0	32.9	17.7	29.4	20.6	14.4	18.4
Tamil Nadu — India	24.2	35.4	17.2	32.1	19.8	15.6	19.8
Miranda — Venezuela	43.0	53.3	30.5	52.5	38.2	38.4	22.8
Romania	42.4	52.4	32.4	55.6	38.3	35.8	23.2
Russia	48.0	56.8	40.5	55.8	43.8	41.9	16.3
Singapore	61.9	67.9	54.4	71.8	58.7	57.8	17.5
Serbia	46.1	57.1	36.4	61.2	44.4	37.9	24.8
Slovak	52.1	60.6	45.8	61.3	48.0	45.2	16.1
Slovenia	49.3	58.7	46.4	58.9	42.3	40.3	18.6
Sweden	54.4	61.8	48.8	63.5	48.3	48.8	15.2
Chinese Taipei	58.2	66.3	48.3	70.7	55.4	52.9	22.3
Thailand	39.4	47.9	31.6	51.0	32.9	33.3	19.4
Trinidad and Tobago	40.8	51.5	32.8	50.1	36.8	33.3	18.8
Tunisia	35.6	44.3	26.9	43.5	31.0	31.0	17.4
Turkey	46.1	54.7	36.0	55.0	39.9	41.8	19.0
Uruguay	42.1	52.3	32.0	52.1	40.5	35.5	20.3
United States	53.8	61.1	46.5	61.4	46.9	49.8	14.9
Average	47.8	56.7	40.1	57.3	43.2	41.8	17.2
Min	23.0	32.9	17.0	29.4	19.8	14.4	12.0
Max	69.3	74.9	57.5	79.4	65.8	68.7	24.9

Note. MC = multiple choice; CMC = complex multiple choice; Closed = closed-constructed response; Short = short response; Open = open-constructed response.

Table A4 Incorrect Rates of Item Types by Country or Region

	Rate (%)						
Country/region	All	MC	CMC	Closed	Short	Open	Max – min
Albania	39.3	39.9	59.0	29.5	34.6	32.2	29.5
UAE	45.5	38.3	59.9	34.6	47.0	46.9	25.3
Argentina	36.4	34.1	54.3	27.5	32.3	32.4	26.8
Australia	35.0	32.0	44.5	27.8	38.6	33.8	16.7
Austria	37.1	35.8	46.2	29.6	37.8	35.4	16.7
Azerbaijan	41.7	46.0	68.1	26.6	28.9	30.4	41.5
Belgium	33.9	30.8	45.6	25.4	36.6	32.5	20.2
Bulgaria	37.0	36.5	56.6	28.9	33.2	30.1	27.7
Brazil	50.2	43.6	63.5	42.0	48.7	51.8	21.5
Canada	35.8	32.1	46.6	28.6	38.1	35.0	18.0
Switzerland	35.5	33.5	45.9	29.4	35.5	33.6	16.5
Chile	40.1	33.9	57.8	32.0	37.3	39.7	25.9
Colombia	43.5	36.3	59.8	34.0	43.1	44.7	25.8
Costa Rica	45.8	36.8	61.5	36.7	45.1	48.8	24.9
Czech Republic	32.9	30.1	45.9	25.0	33.8	30.5	20.9
Germany	33.7	32.0	42.9	28.1	35.3	31.7	14.8
Denmark	37.9	34.9	49.5	30.0	40.5	36.2	19.6
Spain	36.4	33.5	48.8	29.8	36.0	34.4	19.0
Estonia	35.7	33.8	47.3	27.0	37.1	32.9	20.3
Finland	31.4	27.0	40.7	24.1	36.2	31.9	16.7
France	32.5	30.9	45.7	26.1	34.4	28.0	19.6
United Kingdom	37.7	35.6	46.4	30.8	41.7	35.8	15.6
Georgia	38.5	39.3	58.1	30.3	31.2	31.1	27.8
Greece	36.8	35.1	52.4	29.6	40.9	30.9	22.8
Hong Kong — China	31.3	29.2	43.6	21.6	35.7	27.7	22.0
Croatia	40.3	38.8	53.5	30.4	43.0	36.4	23.1
Hungary	36.3	33.0	49.3	31.3	37.7	33.5	18.0
Indonesia	51.9	50.4	65.9	43.0	50.3	48.5	22.8
Ireland	36.5	34.9	46.3	30.5	39.4	33.7	15.8
Iceland	36.6	31.8	49.3	28.0	37.0	36.6	21.2
Israel	36.5	34.2	52.5	30.7	39.8	30.8	21.8
Italy	35.3	32.1	48.5	29.2	37.3	32.6	19.3
Jordan	48.7	41.3	64.7	42.1	50.8	48.4	23.4
Japan	29.6	30.5	41.8	21.0	32.8	23.1	20.7
Kazakhstan	44.6	39.5	60.6	35.4	40.7	44.2	25.2
Kyrgyzstan	40.5	42.5	56.5	33.5	34.0	33.3	23.2
Korea	31.9	26.3	45.1	21.8	37.5	31.8	23.4
Liechtenstein	34.3	32.5	45.1	30.6	34.7	31.3	14.5
Lithuania	41.2	39.3	55.2	30.9	41.5	38.1	24.3
Luxembourg	36.7	35.1	47.8	29.7	37.8	33.8	18.1
Latvia	39.9	36.5	52.2	33.6	43.2	37.6	18.6
Moldova	41.7	43.5	59.6	32.9	36.6	33.8	26.7
Mexico	46.9	37.8	60.5	35.0	44.9	51.9	25.5
Malta	36.6	33.2	52.7	27.0	35.3	34.0	25.7
Montenegro	42.7	44.5	62.5	30.6	40.8	33.6	31.9
Mauritius	46.5	42.7	62.6	37.5	42.5	44.8	25.1
Malaysia	45.3	39.2	63.4	35.5	45.9	43.7	27.9
Netherlands	37.3	31.8	44.7	29.3	39.3	40.6	15.5
Norway	35.2	32.3	46.1	29.9	35.5	33.6	16.3
New Zealand	32.9	29.9	41.6	26.2	37.2	31.8	15.4
Panama	48.5	43.2	63.6	42.5	46.1	47.9	21.1
Peru	43.8	38.5	58.4	33.6	40.1	44.6	24.7
Poland	35.4	32.7	47.4	26.9	39.6	33.0	20.5
Portugal	37.3	35.0	50.3	30.8	38.9	33.8	19.4
Qatar	51.7	48.5	66.2	42.0	50.7	49.4	24.2
Shanghai — China	28.5	24.5	42.7	19.0	30.8	26.6	23.7
Himachal Pradesh — India	51.8	51.4	68.0	42.7	49.1	45.9	25.2

Table A4 Continued

	Rate (%)						
Country/region	All	MC	CMC	Closed	Short	Open	Max – min
Tamil Nadu — India	65.7	58.9	76.7	57.1	65.0	69.0	19.6
Miranda — Venezuela	36.9	31.5	55.5	29.9	35.7	34.3	25.6
Romania	50.2	43.8	63.9	37.6	50.8	51.9	26.3
Russia	37.7	35.2	51.8	31.5	37.8	34.0	20.2
Singapore	32.7	29.2	43.5	24.3	34.0	32.4	19.2
Serbia	39.3	36.2	57.3	27.9	35.2	36.4	29.4
Slovak	37.7	36.1	51.0	30.4	37.6	33.8	20.7
Slovenia	38.9	37.2	50.8	31.6	40.0	35.8	19.2
Sweden	35.0	32.5	45.8	26.9	38.1	32.9	18.9
Chinese Taipei	35.7	31.8	50.1	24.8	35.6	34.3	25.3
Thailand	52.5	47.0	64.4	41.0	54.6	54.2	23.4
Trinidad and Tobago	38.3	33.5	53.6	30.1	37.1	37.0	23.5
Tunisia	47.4	43.4	63.2	39.6	47.2	45.0	23.7
Turkey	44.8	41.4	60.3	36.6	46.5	41.6	23.7
Uruguay	36.5	33.1	54.7	28.2	31.6	33.5	26.5
United States	42.6	37.2	51.6	34.9	46.9	44.2	16.7
Average	39.7	36.5	53.5	31.3	39.8	37.4	22.2
Min	28.5	24.5	40.7	19.0	28.9	23.1	14.5
Max	65.7	58.9	76.7	57.1	65.0	69.0	41.5

Note. MC = multiple choice; CMC = complex multiple choice; Closed = closed-constructed response; Short = short response; Open = open-constructed response.

Table A5 Mean WLE Abilities Under Three Conditions on Item-Level Nonresponses by Country or Region

Country/Region	Condition A	Condition B	Condition C	Condition B – A	Condition C – A
Shanghai — China	1.393	1.434	1.474	0.042	0.081
Korea	1.210	1.253	1.347	0.044	0.138
Hong Kong — China	1.127	1.181	1.275	0.054	0.147
Finland	1.089	1.149	1.264	0.060	0.175
New Zealand	0.987	1.061	1.185	0.075	0.198
Singapore	0.983	1.048	1.146	0.064	0.163
Japan	0.959	1.032	1.276	0.073	0.317
Netherlands	0.910	0.949	0.982	0.039	0.072
Canada	0.848	0.921	1.042	0.073	0.194
Belgium	0.809	0.874	1.027	0.065	0.218
Australia	0.793	0.874	1.004	0.081	0.211
Poland	0.778	0.838	0.976	0.060	0.199
Estonia	0.757	0.818	0.936	0.062	0.180
Norway	0.739	0.820	0.985	0.081	0.246
Iceland	0.701	0.793	0.947	0.092	0.247
Hungary	0.699	0.753	0.922	0.053	0.223
Liechtenstein	0.681	0.752	0.903	0.070	0.222
France	0.681	0.785	1.036	0.104	0.354
Chinese Taipei	0.676	0.732	0.841	0.056	0.165
Sweden	0.675	0.775	0.976	0.100	0.301
United States	0.674	0.729	0.775	0.055	0.101
Ireland	0.666	0.748	0.904	0.082	0.238
Czech Republic	0.662	0.714	0.927	0.052	0.266
Germany	0.646	0.704	0.896	0.058	0.250
Switzerland	0.643	0.703	0.865	0.060	0.222
United Kingdom	0.612	0.673	0.822	0.061	0.210
Italy	0.599	0.690	0.893	0.091	0.294
Portugal	0.575	0.671	0.836	0.096	0.261
Latvia	0.570	0.643	0.744	0.073	0.174
Greece	0.550	0.654	0.881	0.105	0.331

Table A5 Continued

Country/Region	Condition A	Condition B	Condition C	Condition B – A	Condition C – A
Macao — China	0.541	0.628	0.749	0.087	0.208
Spain	0.527	0.620	0.801	0.093	0.274
Denmark	0.472	0.547	0.724	0.075	0.252
Slovak	0.440	0.495	0.694	0.055	0.253
Luxembourg	0.426	0.509	0.715	0.083	0.289
Israel	0.414	0.541	0.808	0.127	0.394
Croatia	0.408	0.458	0.630	0.050	0.222
Austria	0.378	0.431	0.638	0.053	0.260
Lithuania	0.335	0.391	0.517	0.057	0.183
Turkey	0.282	0.345	0.487	0.063	0.205
Russia	0.241	0.372	0.591	0.131	0.350
Chile	0.224	0.315	0.503	0.092	0.280
Malta	0.196	0.270	0.534	0.075	0.338
Slovenia	0.191	0.235	0.461	0.044	0.270
Costa Rica	0.094	0.213	0.324	0.120	0.230
Serbia	0.090	0.126	0.379	0.035	0.289
UAE	−0.003	0.057	0.210	0.059	0.213
Miranda — Venezuela	−0.025	0.167	0.433	0.192	0.458
Mexico	−0.044	0.091	0.180	0.134	0.223
Bulgaria	−0.071	0.006	0.323	0.077	0.393
Colombia	−0.128	0.087	0.223	0.215	0.351
Romania	−0.137	−0.107	0.018	0.030	0.155
Uruguay	−0.154	0.038	0.352	0.192	0.507
Thailand	−0.180	−0.112	−0.019	0.068	0.161
Trinidad and Tobago	−0.221	−0.022	0.283	0.199	0.504
Malaysia	−0.232	−0.169	0.076	0.062	0.307
Jordan	−0.279	−0.213	−0.058	0.066	0.221
Brazil	−0.385	−0.273	−0.126	0.112	0.259
Argentina	−0.430	−0.205	0.150	0.225	0.581
Mauritius	−0.452	−0.392	−0.164	0.060	0.288
Tunisia	−0.452	−0.320	−0.113	0.132	0.340
Montenegro	−0.460	−0.387	−0.019	0.073	0.441
Indonesia	−0.495	−0.369	−0.193	0.127	0.302
Kazakhstan	−0.503	−0.389	−0.154	0.113	0.349
Moldova	−0.571	−0.479	−0.168	0.092	0.403
Albania	−0.614	−0.533	−0.127	0.081	0.487
Panama	−0.715	−0.552	−0.325	0.163	0.390
Georgia	−0.727	−0.562	−0.140	0.165	0.587
Peru	−0.759	−0.540	−0.287	0.219	0.473
Qatar	−0.779	−0.719	−0.479	0.060	0.300
Azerbaijan	−0.884	−0.845	−0.318	0.039	0.567
Tamil Nadu — India	−1.295	−1.304	−1.160	−0.009	0.134
Kyrgyzstan	−1.416	−1.148	−0.731	0.267	0.685
Himachal Pradesh — India	−1.466	−1.406	−1.052	0.060	0.414

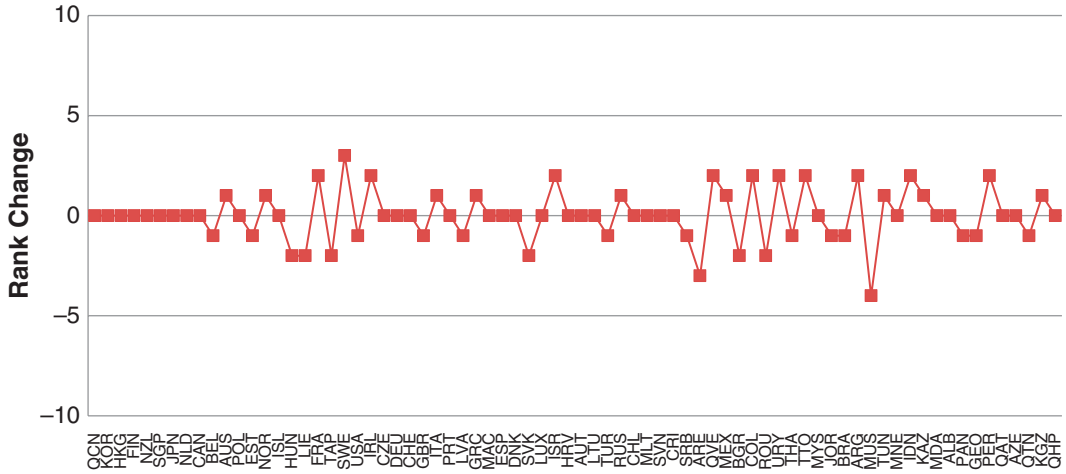

Figure A1 Rank change under Condition B from Condition A for all countries and regions.

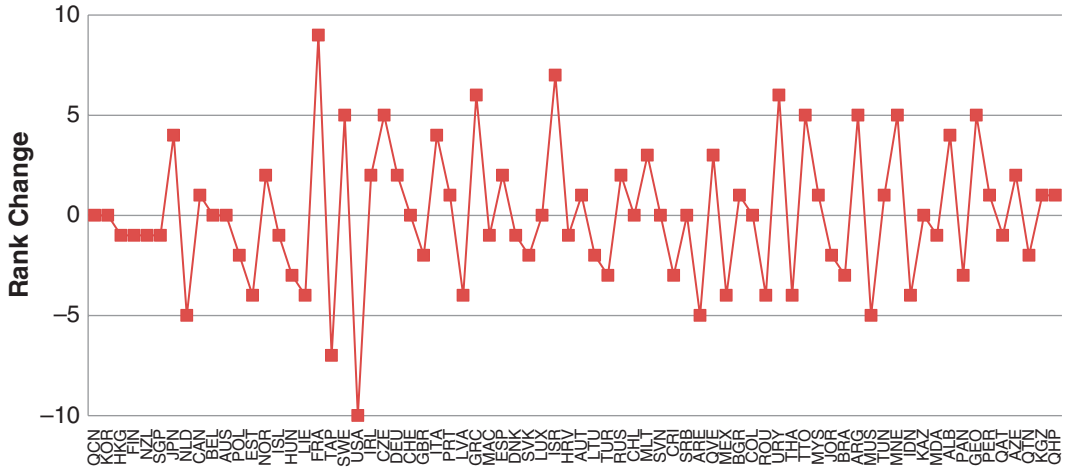

Figure A2 Rank change under Condition C from Condition A for all countries and regions.

Suggested citation:

Chen, H. H., von Davier, M., Yamamoto, K., & Kong, N. (2015). *Comparing data treatments on item-level nonresponse and their effects on data analysis of large-scale assessments: 2009 PISA study* (ETS Research Report No. RR-15-12). Princeton, NJ: Educational Testing Service. 10.1002/ets2.12059

Action Editor: Shelby Haberman

Reviewers: Daniel McCaffrey and Yue Jia

ETS, the ETS logo, and LISTENING. LEARNING. LEADING. are registered trademarks of Educational Testing Service (ETS). All other trademarks are property of their respective owners.

Find other ETS-published reports by searching the ETS ReSEARCHER database at http://search.ets.org/researcher/