Chapter 13
Investigating Rater Effects in International Large-Scale Assessments

Hyo Jeong Shin, Matthias von Davier and Kentaro Yamamoto

Abstract The present study investigates rater effects in international large-scale assessments to evaluate the construct validity of constructed-response items scored by human raters. Using the Programme for International Student Assessment data collected in 2015, we illustrate the methods and present the findings about rater effects on constructed-response items in the context of the first fully computer-based international student skill survey. By comparing the unidimensional and the two-dimensional multiple-group item response theory model, it was shown that the latent correlations between human- and machine-coded items were almost perfect. Country means of human- and machine-coded items were generally similar and their differences were small. Further investigation into individual rater effects mostly resulted in small random loadings, which implied that rater effects were negligible in most countries.

13.1 Introduction

Most cognitive assessments are likely to either include multiple-choice (MC) items, constructed-response (CR) items, or both. Unlike the process for scoring MC items, collecting data from CR items often requires human scoring. In the context of international large-scale assessments (ILSAs), ratings from human coders[1] constitute a significant portion of the data and are an essential factor for data quality and, eventually, the reliability and validity of test results. It is common practice to monitor rater reliability by obtaining multiple scores from different raters for a subset of responses

[1] In the context of PISA, human raters are often called "coders." In this chapter, "coders (coding)" and "raters (rating)" are used interchangeably.

H. J. Shin (✉) · K. Yamamoto
Educational Testing Service, Princeton, NJ, USA
e-mail: hshin@ets.org

M. von Davier
National Board of Medical Examiners, Philadelphia, PA, USA

B. P. Veldkamp and C. Sluijter (eds.), *Theoretical and Practical Advances in Computer-based Educational Measurement*, Methodology of Educational Measurement and Assessment, https://doi.org/10.1007/978-3-030-18480-3_13

and evaluating the level of agreement between multiple human raters. For example, human raters in the Programme for International Student Assessment (PISA) are expected to achieve a minimum of 85% agreement with other raters at the item level, based on the subset of common responses. Within each core domain (Mathematics, Reading, Science), an agreement rate of 92% across items is expected. Once the target level of inter-rater reliability is achieved, human-rated scores are assumed to be valid and reliable, and scores from the lead human raters are then used as the final scores in subsequent analytic procedures (such as item calibrations) along with MC item scores (Organisation for Economic Co-operation and Development [OECD] 2017).

Agreement rates between multiple human raters may be sufficient within a given country for a given assessment cycle to examine whether they understand and apply the scoring rubric consistently. In the context of ILSAs, however, it is imperative to also investigate whether raters from different countries and different cycles provide comparable ratings. The main goal of ILSAs is to compare the skills, knowledge, and behaviors of various populations across countries, focusing on group-level scores (Kirsch et al. 2013). This would not be possible without comparability of human-rated scores across countries and cycles; otherwise, the reliability and validity of the test results would be threatened within and across countries and cycles.

Therefore, for CR items to be used effectively in ILSAs, studies have emphasized that raters should be trained to assure that scoring remains valid and comparable across countries and cycles (Mazzeo and von Davier 2008, 2014). Although concerns regarding the validity of human-rated scores and the potential presence of rater bias have been raised in the research literature, a comprehensive investigation has thus far been almost impossible due to a lack of data recording in paper-based assessments (PBA) from the past. For example, in PISA cycles prior to 2015, only a subset of CR items was examined for reliable agreement rates among raters. Coding PBA items in the online coding system was not compulsory. The process saw most participating countries coding CR paper-based items on multiple paper-based coding sheets associated with test booklets, and then entering data into the data management software (OECD 2014). With the introduction of computer-based assessment (CBA) in the 2015 PISA cycle, however, test takers' raw responses to all CBA CR items and corresponding scores, along with the rater assignments on those scores, have been recorded and are easily accessible. Switching to CBA opened up the possibility of investigating the potential presence of rater bias and the reliability and validity of human-rated scores.

Therefore, this study aims to investigate the effect of human rating on construct validity in ILSAs in terms of rater effects that often have been neglected in the past. We use the 2015 PISA data collected for Science, which was the major domain for this cycle and included over 50 human-rated CR items along with over 100 MC items.[2] Two research questions are investigated in this study: (a) the extent to which MC items and CR items similarly measure the underlying latent construct, and (b)

[2] The major domains for PISA included a large number of both new and trend items. The minor domains included a small number of only trend items.

the extent to which individual rater effects exist in the PISA data across countries. In order to answer these research questions, we first apply and compare unidimensional and multidimensional item response models using appropriate variants of these models for analyzing cognitive tests with mixed item types (i.e., MC items and CR items), with different scoring types (i.e., machine- and human-scored items) while accounting for the clustering (i.e., participating countries). We then fit multiple linear regression models specifying rater indicators as input variables to predict the resultant person proficiencies from the multidimensional item response theory (IRT) models. This allows estimation of individual rater effects that are comparable across countries.

13.2 Scoring Human-Coded Items in PISA 2015

13.2.1 Categorization of Items by Item Formats

The definition of item formats can be different and vary by the purpose of the assessment or the target construct. Following the typology of Bennett et al. (1990), an MC item is defined as any item in which the test taker is required to choose an answer from a relatively small set of response options. A CR item is defined as an item that requires the test taker to compose his or her own answer. In PISA, test takers are given a test form consisting of both MC and CR items organized in groups, or testlets, based on a common stimulus. Historically, five format types were identified and initially used in PISA (OECD 2001)[3]: two types of MC (regular and complex multiple-choice items), and three types of CR (closed constructed response, short response, and open constructed response).

Approximately one third of items across the core domains in PISA 2015 were CR items that required human coding, with nearly 50% of them in the Reading domain being CR. To investigate rater effects across countries in ILSA, we used the PISA 2015 Main Survey data and focused on the Science domain, which, as the major domain in the 2015 cycle, had newly developed frameworks and new items (OECD 2016). With the introduction of the CBA mode in PISA from the 2015 cycle, some of the CR items could be scored by a computer while many of the CR items still required coding by human raters. Items with numeric responses (i.e., only numbers, commas, periods, dashes, and back slashes), responses involving choices from a drop-down menu (i.e., radio-button-type items), or selecting rows of data were scored by computer. All others, typically answered by entering text-based answers, were coded by human raters. Hence, the item format and the traditional type of scoring did not necessarily align, so we used "item formats" to distinguish between MC items

[3] Switching to the CBA mode enabled the development and administration of interactive and innovative item formats, such as simulation-based items. In this study, we focus on the distinction between CR and MC, but readers can refer to the PISA 2015 technical report (OECD 2017) for such new item formats.

Table 13.1 Categorization of CBA science items from the PISA 2015 main survey

	Machine-scored	Human-rated
New	69	30
Trend	57	28
Total	126	58

Table 13.2 Number of human coders per country in PISA 2015 science domain

Number of countries	Number of coders	Sample size
2	4	1501–4000
40	8	4001–7000
4	12	7001–9000
6	16	9001–13,000
5	20	13,001–19,000
1	32	19,001–29,000
1	36	More than 29,000

and CR items, and "scoring types" to distinguish between human-rated items and machine-scored items. Because human-rated items were a subset of CR items and machine-scored items included a small number of CR items and all MC items, the focus of comparison was between "human-rated CR items" versus "machine-scored CR and (all) MC items."

There were 184 CBA Science items analyzed in total: 58 human scored (8 polytomous and 50 dichotomous items), and 126 machine scored (5 polytomous and 121 dichotomous items). Ninety-nine items were newly developed items, and 85 were trend items that had been administered in previous cycles. With about a third of new items (30 of 99) being human rated, the importance of studying the validity of human ratings is clear (Table 13.1)

In analyzing the data, we included countries that administered the test via CBA mode only; information about rater assignment linking to their scores was inaccessible via PBA mode. Overall, 59 CBA countries were included in the study. Table 13.2 presents the number of human coders per country in PISA 2015 Science. According to the coding design and procedures (Chapter 13 of the PISA 2015 technical report; OECD 2017), most countries (40 of 59) followed the standard design with 8 coders (grayed in Table 13.2). Two countries with smaller sample sizes had 4 coders, while the 17 remaining countries with larger sample sizes had 12, 16, 20, 32, and 36 coders, respectively, depending on the increasing size of the sampled language group of the assessment.

13.2.2 Coding Design and Procedures

Human coders who evaluate CR item responses and provide scores must be trained to ensure that they adhere to scoring rules, which should be applied consistently for the given assessment, as well as over multiple assessment cycles and across participating countries (Mazzeo and von Davier 2008). Scoring of CR items by human raters is time consuming, expensive, and can be subject to bias, meaning there is potential for inconsistencies across raters, cycles, or countries. Therefore, coder reliability checks in PISA 2015 were for the first time designed to evaluate within- and cross-country levels for all human-rated CR items, facilitated by a coding design that involved *multiple coding*, or coding of the same response independently by different raters. In PISA 2015, it was assumed that the typical number of raw responses to be coded in a single country-language group was around 180,000. Coding design and procedures were made assuming that a single human coder could code 1000 responses per day (it would take 180 days if a single person were to complete the task alone). Multiple coding of all student responses in an international large-scale assessment like PISA is labor intensive and costly. Thus, a subset of test takers' responses was used to evaluate the inter-rater reliability within and across countries.

Within country, 100 student responses per human-coded item were randomly selected for multiple coding. The rest were evenly split among multiple human coders for single coding. The within-country reliability of raters (i.e., exact agreement rates among human raters based on the multiple coding) varied across items and countries. In PISA 2015, 96% of the CBA countries coded every item with proportion agreement higher than 85% in the new Science items. More than 97% of CBA countries had five or fewer items with proportion agreements lower than 85% in the trend Science items. For most CBA countries, the standard inter-rater reliability concerning exact agreement rate was at least 90% for all domains (90% in new Science, 93% in trend Science).

Coder reliability was also evaluated across countries to ensure that the scores on CR items were comparable across countries. In the PISA 2015 cycle, two coders who were bilingual in both English and the language of the assessment additionally scored 10 "anchor responses" for each item. These responses were written in English and were used to compare the application of scoring rules across countries.[4] In general, across-country agreement tended to be lower than within-country agreement; while a mean of 94.2% within-country agreement was observed for trend and new Science items, slightly lower means of 93.6% for trend Science and 93.1% for new Science items were observed across countries. Since the number of anchor responses was only 10 in PISA 2015 and previous cycles, it was increased to 30 for PISA 2018 to make the comparison more robust.

[4] "Anchor responses" were called "control scripts" up to the PISA 2012 cycle, with the number of control scripts varying between 2 and 19 depending on the items (OECD 2014).

13.3 Construct Equivalence of Different Scoring Types in PISA

Messick (1993) argued that “trait (construct) equivalence is essentially an issue of the construct validity of test interpretation and use (p. 61).” With mixed item formats and diverse scoring types in the test, construct equivalence is critical for using and interpreting combined scores from items in different formats and different scoring types as single trait scores. Studies about construct equivalence mostly have focused on the item format, that is, whether MC items and CR items measure the same latent trait and whether the scores from different formats are comparable and valid. For example, Traub (1993) argued “the true-score scales of the MC and CR instruments must be equivalent for the comparison of difficulty to be meaningful” and concluded that there is too little sound evidence in hand regarding the equivalence of item formats, adding that it would vary by domain. More recently, Rodriguez (2003) reviewed the literature and argued that when items are constructed in both formats using the same stem, the correlation between them is higher than using non-stem equivalent items, concluding that the equivalence appeared to be a function of the item design method or the item writer’s intent. In the context of ILSAs, O’Leary (2001) found that item format could influence countries’ rankings in Trends in International Mathematics and Science Study (TIMSS) data and suggested that different experiences of using MC and CR items in different countries can have even more important implications for interpreting test scores. Hastedt and Sibberns (2005) presented a similar conclusion—that there were differences for subgroups of students and for different countries in achievement by item format on TIMSS.

Essentially, construct equivalence of different item formats is closely related to the different scoring types because mostly human raters score CR items. This notion of construct equivalence of different scoring types—whether machine and human-rated items provide consistent information—has not been thoroughly studied in ILSAs yet. Instead, for practical reasons, the focus has been to achieve the operational standards of inter-rater reliability for the CR subset of item responses. Therefore, to examine construct equivalence of the different scoring types, that is, to see if both machine-scored items and human-rated items measure the same construct, we first estimate the correlations between “human-rated CR items” versus “machine-scored CR and (all) MC items.”

13.3.1 Methods

We fit multiple-group IRT models (Bock and Zimowski 1997; von Davier and Yamamoto 2004) the same way as was performed operationally in PISA 2015 (OECD 2017). Multiple-group IRT models enable the estimation of item parameters that are common across different populations, as well as unique group means and standard deviations. Let j denote a person in group k responding in category x_{ij} of item i, and suppose there are g groups and a test composed of n items. Assuming conditional independence of responses, the probability of observing the pattern of response ($\boldsymbol{x}_j$= [x_{1j}, x_{2j},…, x_{nj}]) can be written as

$$P\left(\boldsymbol{x}_j|\theta\right) = \prod_{i=1}^{n} P_i(X_i = x_{ij}|\theta)$$

which applies to all groups and persons, given the person attribute θ. More precisely, the two-parameter logistic model (Birnbaum 1968) for dichotomous items and the general partial credit model (Muraki 1992) for polytomous items were used in modeling the item response function.

Based on these IRT models, items are characterized by item slopes and item locations (difficulties), and the item parameters can be either constrained to be the same across different groups or allowed to be unique for each group. A latent person ability, or attribute θ, follows a continuous distribution with a finite mean and variance in the population of persons corresponding to group k. With the probability density function denoted as $g_k(\theta)$, the marginal probability of response pattern $\boldsymbol{x}_j$ in group k can be expressed as

$$\overline{P_k}\left(\boldsymbol{x}_j\right) = \int_{-\infty}^{\infty} P\left(\boldsymbol{x}_j|\theta\right)g_k(\theta)d\theta.$$

At the item calibration stage in the PISA 2015 Main Survey, each item was allowed to be either common across countries or unique for a specific country-by-language group, or unique within subsets of multiple country-by-language groups (OECD 2017). In this study, we only consider international item parameters because the purpose of the study is to examine rater effects that are comparable across countries. Any systematic rater effect could introduce country-level differential item functioning, and such systematic effects could be already reflected in the unique item parameters allowed for specific countries. Therefore, during the analyses, we fixed all item parameters to the international common item parameters obtained from the PISA 2015 Main Survey for all countries.

In this study, two types of multiple-group IRT models were fit and the results were compared: a unidimensional model that assumed all items measured the Science domain regardless of the scoring type and a two-dimensional model that separated items by scoring types (the first dimension for "human-rated CR items" and the second dimension for "machine-scored CR and [all] MC items.") For this analysis, we used the *mdltm* software (Khorramdel et al., in press; von Davier 2005), which provides marginal maximum likelihood estimation via the expectation-maximization algorithm (Sundberg 1974, 1976; also Dempster et al. 1977).

13.3.2 Findings

Figures 13.1 and 13.2 show distributions of item slopes and item difficulties that were used for fixing the item parameters by item formats. As noted earlier, there were relatively fewer human-coded items (58 vs. 126) in the data. Human-coded items tended to have higher item slopes (i.e., the proportion of items whose slopes are greater than 1) and appeared more difficult compared to machine-coded items (i.e., the percentage of items whose difficulties are greater than zero). This observation is in line with previous studies in large-scale assessments, such as Dossey et al. (1993), who investigated the item formats in the 1992 mathematics assessment of the National Assessment of Educational Progress (NAEP), and Routitsky and Turner (2003), who studied the item formats in PISA 2003. For NAEP, Dossey and colleagues' analyses showed that the extended CR items were much more difficult than MC items and provided considerably more information per item for more proficient students. In a similar vein, using PISA 2003 Field Trial data, Routitsky and Turner asserted that the MC items appeared easier than CR items on average, and CR items showed higher discrimination (i.e., the correlation between item score and the total score) in general.

Table 13.3 shows the comparison of model fit using Akaike information criterion (AIC; Akaike 1974) and Bayesian information criterion (BIC; Schwarz 1978), which penalizes the number of parameters more strongly than AIC. Both model-fit statistics favored the two-dimensional model for the data, which suggested potential differences by scoring types. However, the difference in model-fit improvement based on the Gilula and Haberman (1994) log-penalty measure was negligible. The unidimensional model reached 99.64% model-fit improvement over the baseline model (independence) compared to the more general two-dimensional model. Hence, it

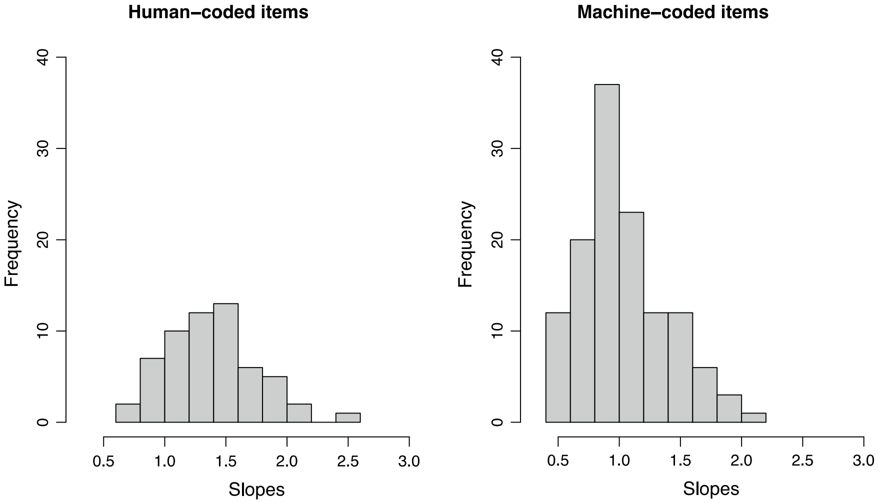

Fig. 13.1 Distribution of item slopes by scoring types (human-coded vs. machine-coded)

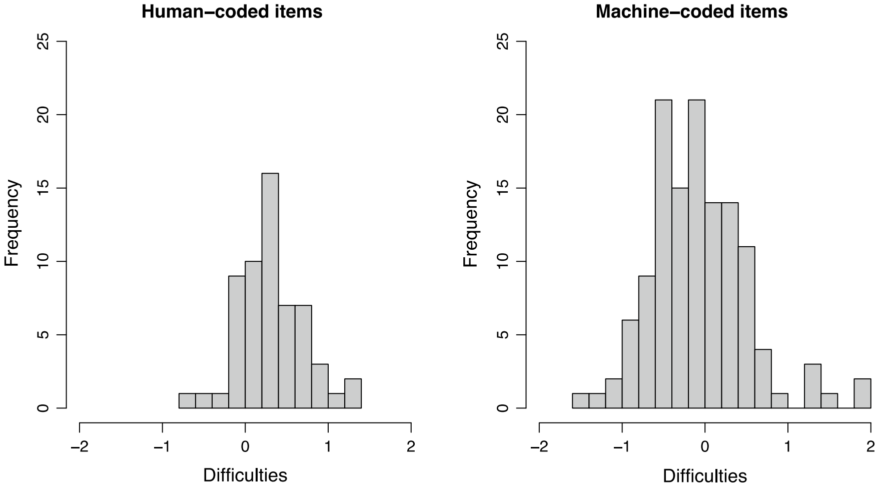

Fig. 13.2 Distribution of item difficulties by scoring types (human-coded vs. machine-coded)

Table 13.3 Comparison of model fit statistics

	AIC	BIC	Log penalty	Percentage improvement
Independence	NA	NA	0.6490	0.00
Unidimensional	9,976,816	9,978,691	0.5668	99.64
Two-dimensional	9,972,179	9,975,302	0.5665	100.00

Note Log Penalty (Gilula and Haberman 1994) provides the negative expected log likelihood per observation; % Improvement compares the log-penalties of the models relative to the difference between the most restrictive and most general model

seems reasonable to assume that MC items and CR items measure a single identifiable latent trait.

Furthermore, the latent correlations were estimated between the two dimensions of scoring types. These correlations were corrected for attenuation by measurement error (Fig. 13.3). The lowest latent correlation among 59 countries was 0.937, and the median latent correlation was 0.974. There were only two countries where the correlation estimates were lower than 0.950. Overall, all countries showed very strong linear associations in their performances on machine- and human-coded items. This implies that human raters may not exert significant effects that should warrant concern about whether human-coded items are considerably different from machine-coded items. Thus, these very high correlations serve as one piece of evidence indicating construct equivalence between different scoring types and the construct validity of the test.

Next, Fig. 13.4 presents the comparison of group-level statistics ($g_k(\theta)$) by different scoring types: the left compares the group (country) means and the right compares the group standard deviations (SD). Regarding the group means, most of

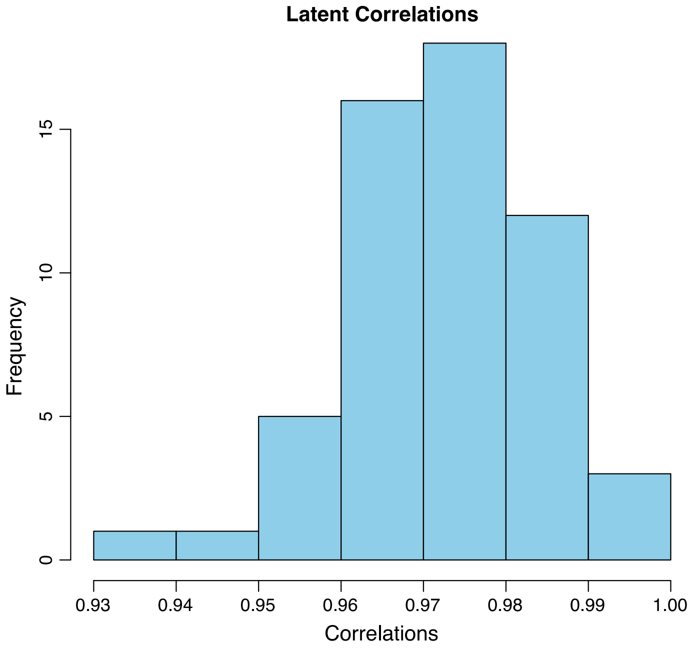

Fig. 13.3 Histogram of latent correlations between machine-coded items and human-coded items

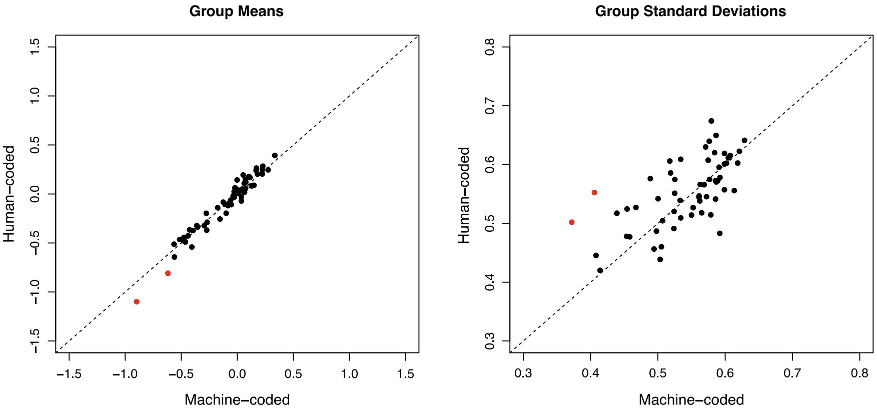

Fig. 13.4 Comparison of group means (left) and standard deviations (right) between machine- and human-coded items

the countries, except for two on the left (solid red dots), showed a consistent pattern in their performance regardless of whether the items were machine or human coded. The shifts in human-coded relative to machine-coded means seem small for all countries, while the largest differences by scoring types were observed in the two lowest-performing countries. Regarding the group SD, there were more notable differences by scoring type, and interestingly, the two lowest performing countries (solid red dots) also showed the largest differences in group SDs: Both showed a higher SD in human-coded items and a smaller SD in machine-coded items. Previous studies noted that the differential behavior of item formats depends on the level of student achievement (DeMars 2000; OECD 2014; Routitsky and Turner 2003). In relation to this, Routitsky and Turner (2003) found a consistent pattern that lower-achieving students performed slightly better on MC than CR items and concluded that lower-achieving countries performed marginally better on MC items because they have a larger proportion of low-achieving students. On the contrary, the OECD (2014, p. 269) calculated the index that indicates achievement on CR items relative to the performance on all other items. The results were that the index was higher in low-performing countries, suggesting that students from low-performing countries were achieving relatively better on the CR items than students from high-achieving countries relative to their achievement on all other items. This pattern might well be attributed to their much lower performance on all other items.

In this study, we suspected that the pattern was observed most likely because these two lowest performing countries could have more missing (omitted) responses on CR relative to MC items. From existing literature it is known that the nonresponse rate is relatively higher for CR items compared to MC items, although that could be the effect of item difficulty or test takers' characteristics (i.e., gender). For example, Routitsky and Turner (2003) reported that CR items tend to comprise difficult sets, with the amount of missing data varying from 9.78 to 57.54% (compared to 1.66–17.62% for MC items), and that item difficulty accounted for about 25% of the missing data (11% for MC items), using PISA 2003 Field Trial data.

Therefore, further investigations were carried out as aimed at skipping response patterns because differences between high and low performers are exaggerated if low performers did not respond to any CR items due to operational scoring rules (e.g., Rose et al. 2017). In present analysis, omitted responses before a valid response were treated as incorrect[5] following scoring rules that have been applied operationally in PISA since its inception. This affects low-performing countries, in particular, as they have omission rates on CR items that are much higher than those observed on MC items relative to other countries. When we calculated the average of the omission rates for each country by scoring types, all 59 country groups showed higher omission rates for human-coded than machine-coded items: The difference in omission rates between human- and machine-coded items ranged between 1.6 and

[5] In PISA, there were two different categories of missing response. If the missing response was followed by valid responses (incorrect or correct) in the subsequent items, it was defined as "omitted." If all the following responses were also missing, it was defined as "not reached." "Omitted" responses were treated as incorrect, but "not reached" responses were treated as missing.

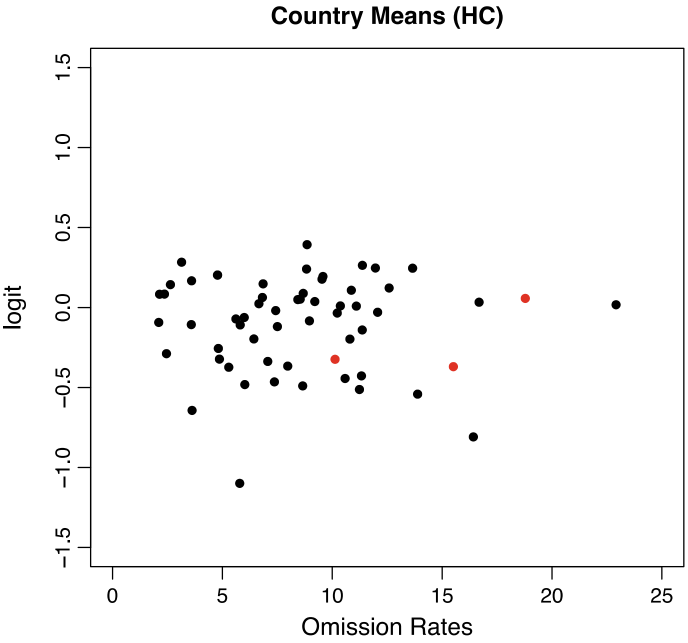

Fig. 13.5 Difference in the country means associated with omission rates

18.4%. Therefore, it was expected that countries with a large number of omits on human-coded items would also tend to have lower scores on those items. Figure 13.5 shows the relationship between those variables when the average omits per country were calculated and associated with the country's performance on the human-coded items. Although the linear association was not seen clearly, three low-performing countries (identified as solid red dots) all showed higher omission rates: Two of them had above 15%, and one had around 10%.

Such a pattern is more apparent when the country means and SDs are sorted by their performance obtained from the unidimensional model that combined two different scoring types. As seen in Fig. 13.6, the difference in the country means by scoring types appears unsubstantial except for the two lowest performing countries. Interestingly, there seems to be a general pattern that lower-performing countries tend to show higher SDs on the human-coded items while higher-performing countries tend to show lower SDs (Figs. 13.6 and 13.7). As shown earlier, this seems to be an artifact due to particularly high nonresponse rates on CR for low-performing countries.

In summary, it seems reasonable to assume that machine- and human-coded items measure the same latent trait, which supports construct equivalence between different scoring types and the construct validity of the test. The two-dimensional model did not add substantial value to that provided by the one-dimensional model, and all countries showed very high correlations of performance measured by human- and machine-

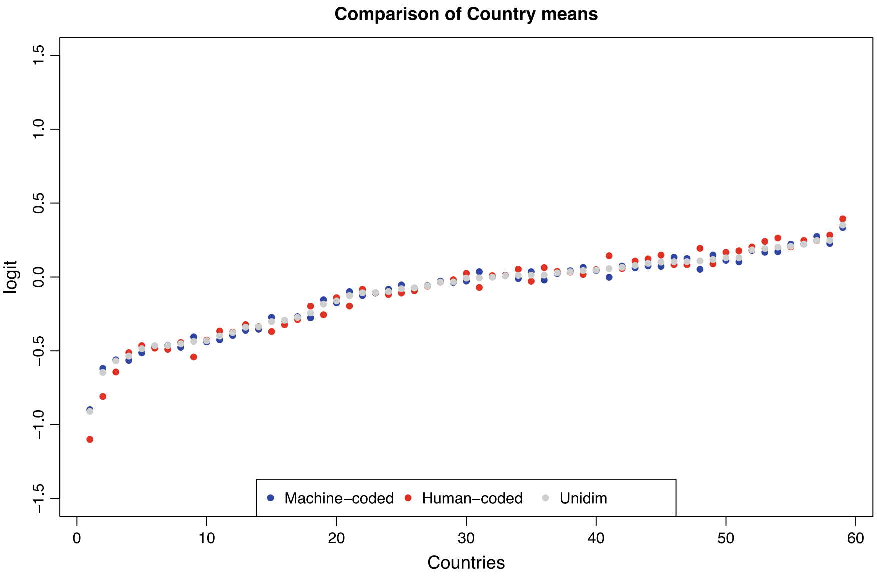

Fig. 13.6 Comparison of country means by scoring types (sorted by group means from the unidimensional IRT model that combined two scoring types)

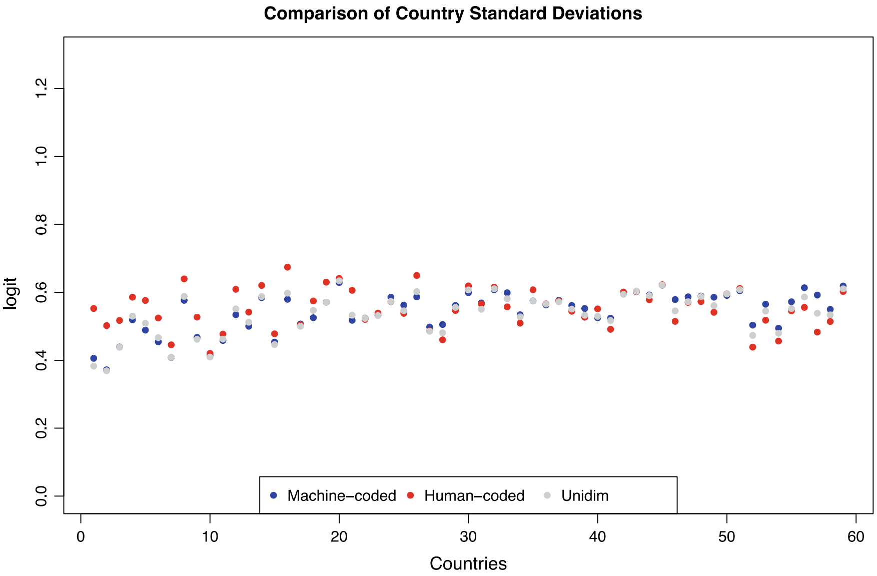

Fig. 13.7 Comparison of country SDs by scoring types (sorted by group means from the unidimensional IRT model that combined two scoring types)

coded items, respectively. Country means of human- and machine-coded items were generally similar and their differences were small for all but the lowest performing countries. The two lowest performing countries showed the largest differences in country means and standard deviations by scoring types, but that pattern seems to be an artifact of scoring rules due to higher omission rates on CR items among low-performing countries. In the next step, we aimed to directly estimate individual rater effects on an internationally comparable scale.

13.4 Rater Effects that Are Comparable Across Countries

13.4.1 Methods

After fitting the two-dimensional multiple-group IRT model separated by scoring types, we fit two types of multiple linear regressions at the international level to estimate the individual rater effects that are comparable across countries. We used the person proficiency estimate as the dependent variable, which was weighted likelihood estimates (WLEs; Warm 1989) based on the second dimension of human-coded items. Below, we use $\widehat{\theta_j^{HC}}$ to denote the WLE of a person j obtained using only the human-coded items and $\widehat{\theta_j^{MC}}$ indicates the WLE for the same person j obtained using only the machine-coded items. To make the interpretation easier, we have standardized those two WLEs to have a mean of 0 and a standard deviation of 1, respectively. In the first model (M1), independent variables included the person proficiency estimates based on the machine-coded items ($\widehat{\theta_j^{MC}}$) and rater-by-country dummy variables. In the second model (M2), only rater-by-country dummy variables were specified to predict the performance on the human-coded items.

$$\widehat{\theta_j^{HC}} = \beta_0 \widehat{\theta_j^{MC}} + \sum_{k=1}^{g} \sum_{r=1}^{R_k} \beta_{rk} N_{rj} d_{rk} \tag{M1}$$

$$\widehat{\theta_j^{HC}} = \sum_{k=1}^{g} \sum_{r=1}^{R_k} \beta_{rk} N_{rj} d_{rk} \tag{M2}$$

Regarding the rater effects, r denotes the rater and R_k the total number of raters in group k. Thus, regression coefficients, β_{rk}, represent the rater effects for rater r in country k, and they are multiplied by N_{rj}, indicating the number of ratings for a person j from rater r associated with the dummy variables (d_{rk}). The number of scores N_{rj} were required because there were multiple raters for each test taker: Different raters were assigned to different item sets and each test taker could have taken a different test form (OECD 2017). In short, each item was coded by a different rater, and each rater was instructed to score a specific set of items. Thus, we calculated the number of item ratings evaluated by individual raters and multiplied those number of ratings to the rater-by-country dummy variables (d_{rk}). These dummy variables take the value 1 for the rater r from the group k, and 0 otherwise. There were 640 dummy variables in total in the data. The number of coefficients is small per country (i.e., no country has more than 36 non-zero dummy variables), while in total we have 640 parameters; per country the number of regression parameters is comparably small. Together with N_{rj} and d_{rk}, the regression coefficients (β_{rk}) represent the average individual rater effects that are weighted by their number of ratings in M2, and weighted average rater effects controlling for the proficiencies on machine-coded items in M1. Finally, the intercept was removed to help the interpretation of the rater effects in a straightforward way, and no country main effects were specified so that rater effects were not centered.

13.4.2 Findings

The overall explained variance at the international level was 0.562 for M1 and 0.135 for M2. A recent study by Shin et al. (in press) analyzed the PISA rater effects of one country and found a negligible effect of raters (4.5% of total variance), arguing that human coding is almost free from subjective bias and that the rater training, scoring rubric, and reliability checks were effective. In this study, the estimate of performance on the machine-coded items ($\widehat{\beta_0}$) in M1 was 0.705 with the standard error of 0.001 (p < 0.001), controlling for the individual rater effects. More importantly, the estimates of individual raters ranged between −0.073 and 0.051 in M1, while the corresponding range was between −0.222 and 0.101 in M2.

Figure 13.8 presents the distributions of 640 individual rater effects ($\widehat{\beta_{rk}}$) that are estimated using an internationally comparable scale across countries. The left panel shows the distribution of estimates obtained from M1 and the right panel from M2. Small random fluctuations in loadings were expected in most countries where rater training worked well. On the other hand, mainly positive loadings were expected in countries where raters tended to be too lenient and negative in countries were expected where raters were too strict. In both cases, the figure clearly shows a majority of small random loadings around zero, which indicated negligible rater effects, hence, rater training seemed to work well in most countries. One notable difference between the two figures was the group of raters with negative loadings (below −0.1) from M2. Nine raters showed severity (below −0.1), and they were all from three low-performing countries, two of which were already identified above in Figs. 13.4 and 13.6 as the lowest performing countries. These raters may have seen many non-responses, particularly in those low-performing countries. Interestingly, when controlling for the performance on machine-coded items, this pattern diminished. Note that M1 controls for ability measured by MC items so that the rater effects were based on differences relative to the conditional expected performance. Thus, the comparison of these figures suggest that the performance of very low-performing students was attributed to raters' severe scoring practice when performance on machine-coded

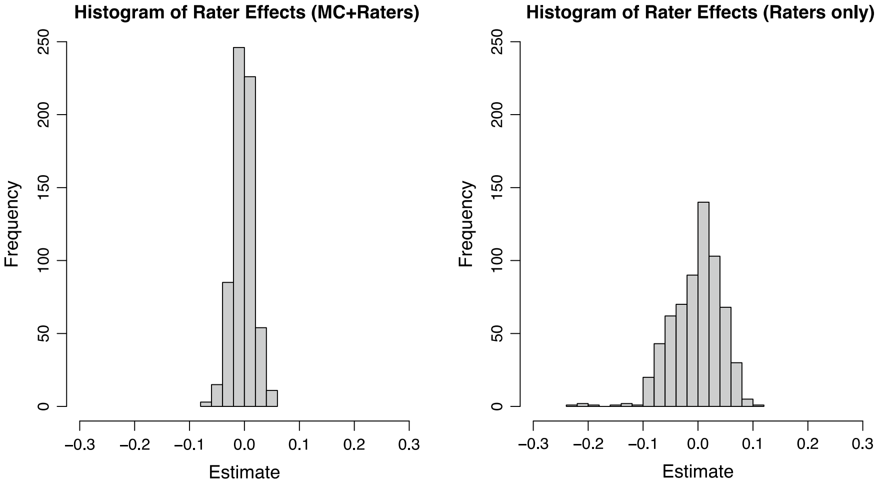

Fig. 13.8 Histogram of rater effects at the international level (left: M1, right: M2)

items was not controlled for. Taken together, small random fluctuations in loadings suggest that rater training worked well in most countries and that human raters did not exert significant effects that should warrant concern about differential rater effects from different countries.

13.5 Conclusion

In the context of ILSAs where a mixed item-format test with a significant portion of CR items is administered, the quality of the scores on the CR items evaluated by human raters is crucial to ensure scoring remains valid and comparable across countries and cycles (Mazzeo and von Davier 2008, 2014). This study investigated rater effects using PISA data collected in 2015 when the assessment switched to CBA as the major mode. Switching from PBA to CBA enabled relatively comprehensive analyses of all CR items, which was previously impossible. By comparing the unidimensional and the two-dimensional multiple-group IRT model, it was shown that the latent correlations between human- and machine-coded items were very high and that the two-dimensional model did not add substantial value to that provided by the one-dimensional model, providing evidence for construct validity by suggesting that the single latent trait is measured by both item types. Country means of human- and machine-coded items were generally similar and their differences were small. Our analysis showed that the two lowest-performing countries also showed the most substantial differences in the country means between two scoring types. Interestingly, larger differences in group SDs were observed as well, and low-performing countries tended to show higher SDs based on the human-coded items compared to the machine-coded items. This seemed to be an artifact of operational scoring rules and higher omitted rates among low-performing countries because high and extremely low performances can be exaggerated if low performers did not respond to any CR items.

Moreover, further investigation of individual rater effects was conducted using multiple linear regressions specifying rater-by-country interactions. The analyses resulted in a distribution of small random loadings, which implied that rater effects were negligible in most countries. There was a group of raters from the three countries whose rater effects were estimated to be too severe, but this effect was diminished when test takers' performance on the machine-coded items were controlled for. In sum, for most countries, rater training seemed to work well, and rater effects were negligible, and hence did not appear to pose any threats to construct validity. At most, two to three countries were identified that might need further investigation on the human-coded CR scores and a review of the scoring rubric and coder training materials. To reiterate, these countries were all low-performing countries where a considerable proportion of students were extremely low achieving, and there seemed no clear evidence of significant rater effects or quality-check efforts specifically on the CR items associated with human coders.

For future studies, rater effects can be investigated concerning the current developments in scoring CR items introduced in ILSAs. For example, the Programme for the International Assessment of Adult Competencies introduced multistage adaptive testing design and a computer-based platform, which automatically and instantaneously scored CR items through the longest common subsequence algorithm (Sukkarieh et al. 2012). More recently in PISA, a machine-supported coding system was implemented operationally for the 2018 cycle and has so far shown to increase the efficiency and accuracy of scoring CR item responses (Yamamoto et al. 2017, 2018). These innovations in scoring CR items using machine learning approaches will reduce the scoring workload of human raters, but it would be worthwhile to investigate whether these tools have a differential effect on scoring correct or incorrect responses, which potentially might affect the estimation of item parameters. A study of the impact of using these tools would be a worthwhile endeavor to assure the measurement invariance and the comparability of item parameters of CR items across countries and cycles.

Finally, systematic rater effects could introduce country-level differential item functioning or indicate the violation of the measurement invariance across countries and cycles. In this study, we have used the data from one assessment cycle and fixed the item parameters using the international parameters obtained at the IRT calibration stage. This was performed intentionally to estimate and separate rater effects that are internationally comparable. A more in-depth look into the comparability across cycles by using multiple cycles of data would be beneficial, and item fit statistics associated with the rater effects might better help the understanding of country- and cycle-level rating processes and behaviors.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*(6), 716–723. https://doi.org/10.1109/tac.1974.1100705.

Bennett, R. E., Ward, W. C., Rock, D. A., & LaHart, C. (1990). *Toward a framework for constructed-response items* (Research Report No. RR–90–7). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1990.tb01348.x.

Birnbaum, A. (1968). Some latent trait models and their use in inferring a student's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

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

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*(1), 55–77. https://psycnet.apa.org/doi/10.1207/s15324818ame1301_3.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B, 39*(1), 1–38.

Dossey, J. A., Mullis, I. V. S., & Jones, C O. (1993). *Can students do mathematical problem-solving? Results from constructed-response questions in NAEP's 1992 mathematics assessment*. Washington, DC: National Center for Education Statistics. https://eric.ed.gov/?id=ED362539.

Gilula, Z., & Haberman, S. J. (1994). Conditional log-linear models for analyzing categorical panel data. *Journal of the American Statistical Association, 89*(426), 645–656.

Hastedt, D., & Sibberns, H. (2005). Differences between multiple choice items and constructed response items in the IEA TIMSS surveys. *Studies in Educational Evaluation, 31*, 145–161. https://eric.ed.gov/?id=EJ723967.

Kirsch, I., Lennon, M., von Davier, M., Gonzalez, E., & Yamamoto, K. (2013). On the growing importance of international large-scale assessments. In M. von Davier, E. Gonzalez & I. Kirsch (Eds.), *The role of international large-scale assessments: Perspectives from technology, economy, and educational research* (pp. 1–11). https://doi.org/10.1007/978-94-007-4629-9.

Khorramdel, L., Shin, H., & von Davier, M. (in press). mdltm (including parallel EM). In M. von Davier & Y. S. Lee (Eds.), *Handbook of diagnostic classification models—State of the art in modeling, estimation, and applications*.

Mazzeo, J., & von Davier, M. (2008). *Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results*. doc.ref. EDU/PISA/GB(2008)28. https://www.researchgate.net/publication/257822388_Review_of_the_Programme_for_International_Student_Assessment_PISA_test_design_Recommendations_for_fostering_stability_in_assessment_results.

Mazzeo, J., & von Davier, M. (2014). Linking scales in international large-scale assessments. In L. Rutkowski, M. von Davier & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 229–258).

Messick, S. (1993). Trait equivalence as construct validity of score interpretation across multiple methods of measurement. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 61–74). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. https://psycnet.apa.org/record/1993-97248-004.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*(2), 159–177. https://doi.org/10.1002/j.2333-8504.1992.tb01436.x.

O'Leary, M. (2001). *Item format as a factor affecting the relative standing of countries in the Trends in International Mathematics and Science Study (TIMSS)*. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.

Organisation for Economic Co-operation and Development. (2001). *Knowledge and skills for life: First results from PISA 2000*. Paris, France: OECD Publishing. http://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/knowledgeandskillsforlifefirstresultsfrompisa2000-publications2000.htm.

Organisation for Economic Co-operation and Development. (2014). *PISA 2012 technical report*. Paris, France: OECD Publishing. https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf.

Organisation for Economic Co-operation and Development. (2016). *PISA 2015 assessment and analytical framework: Science, reading, math, and financial literacy*. Paris, France: OECD Publishing. http://www.oecd.org/publications/pisa-2015-assessment-and-analytical-framework-9789264281820-en.htm.

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

Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement, 40*(2), 163–184. https://doi.org/10.1111/j.1745-3984.2003.tb01102.x.

Rose, N., von Davier, M., & Nagengast, B. (2017). Modeling omitted and not-reached items in IRT models. *Psychometrika, 82,* 795–819. https://doi.org/10.1007/s11336-016-9544-7.

Routitsky, A., & Turner, R. (2003). *Item format types and their influence on cross-national comparisons of student performance*. Presentation given to the Annual Meeting of the American Educational Research Association (AERA) in Chicago, IL. Retrieved from http://works.bepress.com/cgi/viewcontent.cgi?article=1013&context=alla_routitsky.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6,* 461–464.

Shin, H., Rabe-Hesketh, S. & Wilson, M. (in press). Trifactor models for multiple-ratings data. *Multivariate Behavioral Research*. https://doi.org/10.1080/00273171.2018.1530091.

Sundberg, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scandinavian Journal of Statistics, 1*(2), 49–58.

Sundberg, R. (1976). An iterative method for solution of the likelihood equations for incomplete data from exponential families. *Communications in Statistics—Simulation and Computation. 5*(1), 55–64. https://doi.org/10.1080/03610917608812007.

Sukkarieh, J., von Davier, M. & Yamamoto, K. (2012). *From biology to education: Scoring and clustering multilingual text sequences and other sequential tasks* (Research Report No. RR –12–25). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02307.x.

Traub, R. E. (1993). On the equivalence of the traits assessed by multiple-choice and constructed-response tests. In R. E. Bennett & W. C. Ward (Eds.), *Construction versus choice in cognitive measurement* (pp. 29-44). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc. https://psycnet.apa.org/record/1993-97248-004.

von Davier, M. (2005). *mdltm: Software for the general diagnostic model and for estimating mixtures of multidimensional discrete latent traits models (Computer software)*. Princeton, NJ: Educational Testing Service.

von Davier, M., & Yamamoto, K. (2004). Partially observed mixtures of IRT models: An extension of the generalized partial credit model. *Applied Psychological Measuremen*t, *28*(6), 389–406. https://www.ets.org/Media/Research/pdf/RR-03-22-vonDavier.pdf.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54,* 427–450. https://doi.org/10.1007/BF02294627.

Yamamoto, K., He, Q., Shin, H., & von Davier, M. (2017). *Developing a machine-supported coding system for constructed-response items in PISA* (Research Report No. RR-17-47). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12169.

Yamamoto, K., He, Q., Shin, H., & von Davier, M. (2018). Development and implementation of a machine-supported coding system for constructed-response items in PISA. *Psychological Test and Assessment Modeling, 60*(2), 145–164. https://doi.org/10.1002/ets2.12169.