Research Report
ETS RR–17-47

Developing a Machine-Supported Coding System for Constructed-Response Items in PISA

Kentaro Yamamoto

Qiwei He

Hyo Jeong Shin

Matthias von Davier

December 2017

RESEARCH REPORT

Developing a Machine-Supported Coding System for Constructed-Response Items in PISA

Kentaro Yamamoto, Qiwei He, Hyo Jeong Shin, & Matthias von Davier

Educational Testing Service, Princeton, NJ

Approximately a third of the Programme for International Student Assessment (PISA) items in the core domains (math, reading, and science) are constructed-response items and require human coding (scoring). This process is time-consuming, expensive, and prone to error as often (a) humans code inconsistently, and (b) coding reliability in international large-scale assessments varies across items and countries. The shift in PISA 2015 from paper- to computer-based assessment digitized all responses and associated coding, thus providing opportunities to introduce technology and analytical methods to improve data processing and analyses in future cycles. In this computer-based setting, this research explains the framework and approach for improving the accuracy and efficiency of the coding process in constructed-response items for future PISA cycles. In the pilot study associated with this research, we demonstrate how the proposed machine-supported coding system can be utilized using sample reading items from the PISA 2015 cycle.

Keywords Machine-supported coding; constructed-response items; human coding; large-scale assessments; PISA

doi:10.1002/ets2.12169

The move toward computer-based assessment (CBA) holds out the promise of great improvements in the quality of data, leading to greater precision and increased validity (e.g., von Davier, Gonzalez, Kirsch, & Yamamoto, 2012). Specifically, it has made greater efficiency in scoring possible, has the potential to make assessment more effective by increasing engagement, and facilitates administration of dynamic and interactive problems. Finally, CBA allows capturing more information about the course of the problem-solving process and reducing the human workload in both administration and the scoring process.

The Programme for International Student Assessment (PISA) is one of the largest internationally standardized assessments, aiming to evaluate education systems worldwide by testing the skills and knowledge of 15-year-old students. To date, students representing more than 70 economies have participated in this triennial assessment that focuses on assessing young people's capacity to demonstrate their preparedness in various domains with a focus on reading, mathematics, and science. The core (or major) domain rotates by cycle. In the PISA 2015 cycle, the major domain was science and was administered to all students, while the minor domains reading and mathematics each were administered to about a third of the students.

For the first time, PISA 2015 delivered the assessments of all subjects via computer. Paper-based assessment instruments consisting only of trend items (i.e., items from previous assessment cycles) were provided for countries that chose not to test their students by computer (Organisation for Economic Co-operation and Development [OECD], 2016). The shift from paper- to computer-based assessment in PISA 2015 led to a fully computer-based data collection and allowed the associated coding process also to be technology-driven.

In this study, we define coding as a process that initially categorizes written responses into a number of discrete classes, thus facilitating the scoring in a later step. The ability to collect students' raw responses and possibly automate the coding of more complex response types — such as extended, constructed answers — is expected to dramatically enhance PISA's overall data quality. In this research report, we outline an approach for improving the accuracy and efficiency of the coding process for future cycles. In particular, we examine the regularities and commonalities of the observed students' raw responses and develop a machine-supported coding system that can be implemented in PISA 2018.

The goal is to avoid repeated coding of the exact same response string by classifying constructed responses into equivalent response classes. (As detailed in Pilot Study section of this report, we grouped only by identical responses, but in the

Corresponding author: K. Yamamoto, E-mail: kyamamoto@ets.org

future we foresee using broader classes of equivalent responses, i.e., responses with small variations in spelling or capitalization or different amounts of white space.) For each item, these response classes can be determined by populating them from previous data collections as well as current response samples. Each response class has a human-scored response attached to it, meaning that human scoring of previously seen equivalent responses is used to determine whether the response class represents a correct or incorrect response. For response classes with perfect or almost perfect agreement, or with expert verified scores, the score associated with the response class can then be applied to future observations of the same exact response, that is, responses from the same equivalent response class.

This approach parallels automated scoring in the sense that a scoring model is first trained on existing data and then applied to future data. However, unlike commonly used automated scoring processes that generally involve algorithms and the selection of certain text features such as sentence or word length, and machine learning (e.g., *c-rater*® scoring engine, Leacock & Chodorow, 2003; *e-rater*® scoring engine, Burstein, 2003), the proposed method in this paper relies on human scores and exact matching of previously established classes of equivalent responses and newly observed student responses. This means that no computer-based classifications or threshold approach are needed in our current work to determine scores; only exactly matching responses receive a score that was previously established based on human coders. Such an exact matching rule could be easily applied to any language that is used in the international large-scale assessment such as PISA, while commonly used automated scoring models are typically developed for one specific language.

Coding System in PISA 2015

In PISA, test takers are given a mixture of constructed-response and multiple-choice items organized in groups based on a common stimulus. The stimuli often take the form of a reading passage, or a passage combined with some figures, with contents taken from situations deemed culturally appropriate. Approximately a third of the items in the core domains (i.e., math, reading, and science) in PISA 2015 are constructed response and require human coding.[1] The proportion is highest in reading, where nearly 50% of items are constructed-response and require human coding.

Human coding of constructed responses is time consuming, expensive, and prone to error due to a lack of consistency among human coders, which results in low coding reliability. Table 1 presents the average number of human coders required for different domains per country/language groups. Because the focus for PISA 2015 was science, more items were administered in this 1-hour domain and all students had to take it. As such, the number of human coders was higher in science than for the minor domains.

In PISA 2015, typically, the number of raw responses to be coded in a single country per language was around 180,000. Assuming that 1,000 responses can be coded by a single human coder per day, it would take 180 days for a single person to complete the task. The challenge is expected to be greater in PISA 2018 for two reasons: The major domain will be reading, which is more heavily text-based and utilizes a higher proportion of constructed-response items, and more countries are expected to participate.

Coder reliability in PISA 2015 was evaluated at the within- and cross-country levels for all items, which was enabled by a coding design that involved *multiple coding*, or coding of the same response by different individuals. In general, each country needed to randomly select 100 student responses per human-coded item for multiple coding. The rest of the student responses needed to be evenly split among multiple human coders for single coding. Multiple coding of all student responses in an international large-scale assessment like PISA is labor-intensive and costly. The inconsistency of coders varied across items and countries. In PISA 2015, in terms of the student responses, 96% of the CBA countries coded every item with proportion agreement higher than 85% in the mathematics, new science items (new science), and financial literacy. More than 97% of CBA countries had five or fewer items with proportion agreement lower than 85% in the reading and trend science (items from previous cycles) domains; for further detail, see the PISA 2015 Technical Report (OECD, 2017). For the majority of CBA countries (Batch 1 in PISA 2015 data collection), the Cohen's kappa agreement was above 90% for all the domains (97% in mathematics, 90% in reading, 90% in new science, 93% in trend science, and 92% in financial literacy).[2]

Motivation for Developing a Machine-Supported Coding System in PISA

In PISA 2015, approximately 75% of countries chose the CBA mode, which resulted in a large number of digitally recorded constructed responses. This very large database can be used as input for a system that structures responses into uniquely

Table 1 Average Number of Human Coders by Domain per Country/Language in PISA 2015

Domain	Number of human coders		
	Minor	Standard	Large
Mathematics	4	6	6
Reading	6	9	9
Science	8	12	16
Financial literacy	4	4	4

Notes. The number of human coders depended on the domain and sample size of the country/language group, which was defined as minor, standard, and large sample groups. The sample sizes in the minor, standard, and large sample groups range were under 7,000, 7,000 to 9,000, and more than 9,000, respectively.

observed answers and answers given by multiple respondents. A machine-supported coding process that utilizes this structured database in the coding process, therefore, facilitates the goal of reducing countries' coding burden while improving the quality of coded response data. Reducing human coders' workload for coding constructed responses is a primary goal because currently students' raw responses are generally allocated to human coders without any screening: Every single response is included in the coding process, regardless of whether the responses are unique (e.g., "white" vs. "red") or similar (e.g., "30 minutes" vs. "30 min") or even none at all or invalid (such as special characters, e.g., "%^&*,").

We can categorize the raw responses into two types: (a) nonresponse, and (b) identical but having a response. This process allows us to place responses in identical response classes. Determining identical responses—for example, 500 instances of "Region A"—can be used to combine the respondents sharing these responses into a single response class. A *response class* for a particular item is defined as the set of respondents who provided the exact same response (i.e., an identical string of characters) to the same item. Combining 500 identical strings into the same response class would eliminate 499 instances of repetitious coding (or 99.8% for this particular response class). The proportion of workload reduction would decrease differently for each item, depending on the level of response complexity and the relative size of the response classes found in the data. For instance, straightforward responses to short constructed-response items such as "45 miles" as the response to a question about finding a distance between two points will more likely result in larger identical response classes and, hence, lead to a larger workload reduction than moderately complex responses, such as explanations of how a drug functions. More detailed illustrations will be given in the Pilot Study section of this paper.

A second motivation for developing a computer-supported coding system comes from the goal to increase coder reliability and accuracy. As noted before, it is not feasible to double code every response. And even if double coding were possible, while the rate of coding error would be reduced, coding errors could still potentially occur and reduce the reliability of the human-coded responses. Bias could even be introduced if some coders produce results that are systematically more lenient or severe. Unlike humans, a computer-based coding system does not show effects from fatigue or introduce random or systematic errors, and accuracy can be guaranteed once the coding algorithm is fully tested and coding standards are implemented correctly. The variability of accuracy of human coding can manifest over multiple cycles due to variability in coder training. In each cycle, the timing, experience of participants, and monitoring of training may all vary to some extent, all of which contributes to a potential reduction in comparability of the coding process.

Lastly, computer-supported coding can reduce cost and increase efficiency. In a machine-supported coding system, scores can be generated and delivered in real time or asynchronously, typically much faster than with human coding. In addition, the training of human coders could utilize the data from computer-based coding, thus being made more efficient by extracting unique responses and, potentially, extracting particular response features by different performance group (e.g., correct/partial correct/ incorrect) for training purposes.

The remainder of this report is structured as follows: In the next section, we introduce our recommendation for a machine-supported coding system that potentially can be used in PISA 2018 and similar large-scale assessments, providing an overview of the system and workflow. The "Pilot Study" section presents results of a pilot study where the proposed machine-supported coding system was applied. In the final section, we discuss the possibility of expanding the method to a broader context and describe ideas for additional gains in efficiency. Some challenges related to machine-supported coding systems are also discussed.

Developing a Machine-Supported Coding System in PISA

The main idea in developing a machine-supported coding system is to focus only on unique responses for coding. This is facilitated by defining what we call *equivalent response classes* (in this study, we incorporate only identical responses in these classes). The only association of equivalent response classes is via the scored response that was determined based on a number of human coders seeing and scoring the response in previous data collections. These classes represent the subset of respondents who produced exactly the same response. Identical responses should receive the same code or scored response, no matter who has given the response, and no matter who coded the response previously, which is a basic requirement of test fairness regardless of whether the coding is performed by humans or machines.

Hence, with identical responses appearing in these equivalent response classes, the classes can be understood as the group of respondents for which the code received for that particular response is interchangeable. Therefore, once the correct code for a response associated with a response class is known, no further coding effort is required. The same code can thus be applied whenever the particular response is observed again, reducing repetitive coding work. By defining equivalent response classes that give structure to the whole pool of observed responses, we obtain a clustering of respondents into groups that should receive the same codes for the same responses. The relative sizes of response classes, as well as the ratio of the number of unique responses to the total number of responses, provides a simple yet effective measure of the expected gain that can be expected from presenting only unique responses for coding.

In terms of practical implementation, the computer-assisted coding starts by collecting existing data. For example, in PISA 2015, raw constructed-response data from the field and main tests can be collected. These raw responses are tagged with the codes generated by human raters. The respondents are then sorted according to their raw responses in order to derive equivalent response classes, and the codes received within these response classes are checked for consistency. As observed in the current study, the majority of human coders tend to provide the same code for the same response as expected, with very few exceptions.

Once the correct (most prevalent) code is determined for a response class by expert coders, based on a thorough validity and reliability check, applying this code automatically to new responses is then a straightforward algorithmic lookup procedure. The process can be applied in future test administrations using the same items. In our pilot study, we looked at data from PISA 2015, with the results applicable to future administrations such as PISA 2018. When new items are developed for successive cycles, it might be possible to extract equivalent response classes from the pilot and field trial data collection and then complete the classification with the main test data by compiling additional equivalent response classes on the fly.

Figure 1 presents a flowchart that describes the machine-supported coding system for responses to computer administered constructed-response items. The workflow can be divided into two phases: (a) learning from past responses and response codes and (b) applying the learned coding based on equivalent response classes in the machine-supported coding process. In the first phase, historical data — for example the coded raw responses from PISA 2015 (or even data from earlier cycles) — are analyzed. A simple algorithm sorts the raw responses and provides sets of respondents with identical responses (the equivalent response classes). The differences between classes in terms of raw responses define the set of unique responses in the input data. If there is a common response code that applies to all of the respondents sharing the same response, a coded unique response (CUR) pool can be generated based on the equivalent classes that each have a single (or a most prevalent) response code. In the application, or the second phase, machine-supported coding is applied to new uncoded responses: If a new respondent's answer to a constructed-response item is found in the CUR pool for that item, the stored response code is directly applied to this new respondent's answer. Nonresponses can be discarded or scored based on the code to which all previously observed nonresponses were assigned. Only those responses that cannot be matched to an identical response in the CUR pool will be assigned to (multiple) human coders. If multiple coders all agree in terms of the assigned response, it is possible to add this response and the associated response code to the CUR.

A machine-supported coding system of the proposed type can be used on the fly as well. It has the potential to work for completely new constructed-response items without any historical data. More specifically, an algorithm that works on the fly can evaluate whether an observed response is part of the CUR, even if the CUR is initially a nonresponse. Any response not in the CUR will be a new one and be presented to a human coder. Any new, unique response is then given to multiple coders (typically more than two) to assess coder agreement in a larger group of human coders before a response is entered into the CUR.
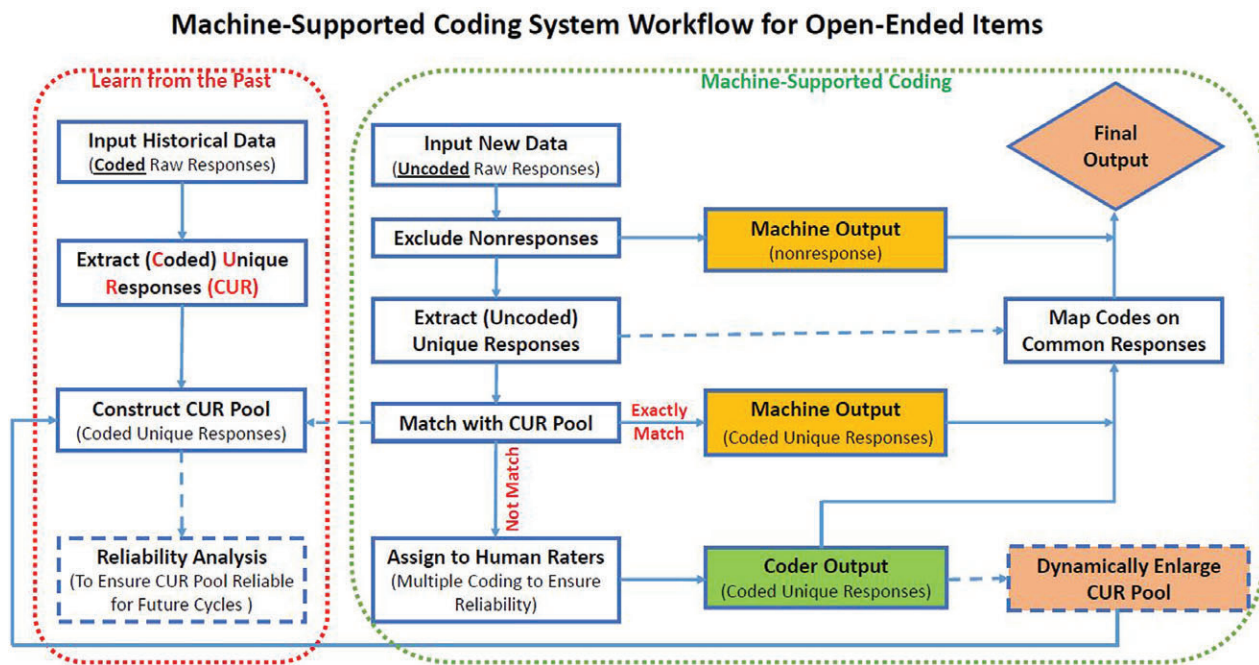
Machine-Supported Coding System Workflow for Open-Ended Items

Figure 1 Machine-supported coding system workflow for constructed-response items.

Existing coding systems can be easily adapted for this purpose. For example, if there are 1,000 raw responses that can be clustered into 100 equivalent response classes, the CUR that can be derived would include 100 unique responses. In this case, only 100 responses need to be coded. If there were four coders, they all could be assigned to coding those 100 unique responses (to ensure reliability) rather than being forced to divide their attention on the larger set of 1,000 raw responses (e.g., two coders independently coding 500 raw responses and two other coders independently coding the other 500). With such a system, human coders' workload can be substantially reduced. Human coding will take place only when necessary, leaving repetitive, operational tasks for the computer.

It is imperative to apply this approach only to responses that are identical to those found in the CUR. This ensures, and indeed enhances, coding fairness, as it provides exactly the same code to all identical responses no matter which respondent provided the response. Human coding error can be reduced, assuming that the multiple coding of a smaller number of unique responses is free of response bias and produces independent coding of the same unique response and that only those responses enter the CUR that are uniformly assigned the same code by all of the coders who saw the response.

At this stage, the system merely optimizes the procedure of assigning coding tasks with the help of technology, which is straightforward and easily applicable in operational settings. Although this strategy is rather simple, it is expected to result in increased efficiency through a reduction of human coders' workload by an average of 40–50%.

Pilot Study: Machine-Supported Coding System in PISA 2015

A pilot study was conducted with the aims of evaluating the feasibility of our proposed machine-supported coding system and examining the cost-effectiveness of this new coding approach. More specifically, we focused on examining the regularities and commonalities of the observed raw responses using 13 example items in the reading domain from the PISA 2015 main survey. We reviewed all 44 constructed-response items used in the reading domain in PISA 2015 and selected the 13 items that varied in the level of challenges that the machine-supported coding may be confronted with (i.e., easy or hard to score). In this pilot study, we examined the extent to which regularities or commonalities are observed from the students' raw responses on constructed-response items in PISA in terms of (a) the pattern and frequency of unique responses and (b) visual representations.

Table 2 Sample Size per Item by Country and Coder Reliability Used in the Pilot Study

	A	B	C	D	E	F	G	Avg. coder reliability (SD)
Item1	1,746	1,215	1,167	732	797	687	644	98.08 (2.06)
Item2	1,815	1,260	1,179	773	837	711	670	99.86 (0.26)
Item3	1,838	1,261	1,188	787	847	710	670	99.81 (0.24)
Item4	1,838	1,261	1,188	787	847	710	670	99.66 (0.55)
Item5	1,838	1,261	1,188	787	847	710	670	99.69 (0.31)
Item6	1,838	1,261	1,188	787	847	710	670	99.71 (0.28)
Item7	1,793	1,212	1,174	747	814	677	647	99.50 (0.63)
Item8	1,789	1,212	1,171	747	813	676	647	99.74 (0.27)
Item9	1,754	1,195	1,148	729	761	690	639	99.86 (0.26)
Item10	1,760	1,195	1,148	732	764	690	639	98.90 (0.91)
Item11	1,782	1,232	1,168	762	827	697	667	97.03 (2.20)
Item12	1,724	1,188	1,333	722	788	690	752	99.65 (0.49)
Item13	1,770	1,198	1,151	736	773	691	639	96.94 (2.94)

Note. The shaded items (Item 2, Item 3, and Item 11) are used to illustrate how the machine-supported coding can be expected to work in the following section. Australia (English), B-S-J-G (China) (Chinese), Germany (German), France (French), Japan (Japanese), Korea (Korean), and the Netherlands (Dutch) are represented by A to G in random order.

Dataset

We examined 13 items in the reading domain based on data from seven countries/regions — Australia (AUS; English), B-S-J-G (China) (CHN; Chinese), France (FRA; French), Germany (DEU; German), Japan (JPN; Japanese), Korea (KOR; Korean), and the Netherlands (NLD; Dutch). This set was selected mainly due to their languages' diversity with, for instance, alphabetic- and character-based languages represented. Also, we chose countries that do not fall into the bottom third of performance as we wanted to avoid the larger proportion of missing data (i.e., nonresponses) that would be expected from low-performing countries. We aimed to compare the pattern of unique responses across countries in order to examine the consistency of unique response distributions across countries and to establish the feasibility of a machine-supported coding system for multiple languages in PISA.

The sample sizes per item by country and average coder reliability used in the pilot study are given in Table 2. In accordance with policies regarding confidentiality and item disclosure, we anonymized all the countries' names, using instead A to G to represent the seven countries in random order. Note that there are slight differences in the sample sizes across items in a given country and that each country has different sample sizes due to sampling requirements, country's population size, and assessment design in PISA 2015 (OECD, 2013). The last column in Table 2 shows the average coder reliability per item in this pilot study, which is fairly high at around 97–98%. Note that the standard deviation of coders' reliability for some items (e.g., Item 11 and Item 13) appears higher than others, suggesting language effects may exist, especially when the set of constructed responses is more diverse. We use three items (Item 2, Item 3, and Item 11, shaded in Table 2) to illustrate how the machine-supported coding can be expected to work and provide more details on unique response distributions and other descriptive measures for these three items in the next section.

Methods

Our pilot study was implemented in three steps: unique response extraction, evaluation of agreement of codes for unique response classes, and comparisons of unique response distributions across countries. In the first step, equivalent response classes and unique responses were extracted from the raw responses separately for each country based on the rule of exact matching for each constructed-response item. The exact matching criterion indicates that no change is made to the raw responses, that is, no implementation of any preprocessing of the raw response. To establish empirical distributions of unique responses, we calculated the frequency of each unique response in the response pool, the frequency of scores coded by human raters (e.g., full credit or no credit[3]), and the frequency of nonresponses. Note that nonresponses can be viewed as one of the equivalent response classes, with the nonresponse constituting one category of the unique response set.

The evaluation of coder agreement for each of the equivalent response classes was undertaken in the second step. Using the unique response sets per item by country, we examined the distribution of assigned response codes within the set of respondents with identical responses, separately for each language. As each coauthor is multilingual with different

language backgrounds, each was able to review two or more datasets used in the study. This is an essential step to examine whether the observed responses were correctly identified by the human coders as either correct or incorrect responses. Very few cases were spotted that indicated miscoding for easy-to-code items, but more miscodes or inconsistent cases were observed for difficult-to-code items.

This analysis provided information on how to revise the coding rubrics and coder training material based on real responses from students. More specifically, by calculating the frequencies of unique responses by full and no credit, we were able to easily identify cases where miscodes were assigned (e.g., full credit was assigned although it should have received no credit based on the coding guides, or vice versa) or where human coders did not agree well (e.g., the same unique response received full-credit codes from some coders but no-credit codes from other coders). Because all the unique responses are based on observed responses generated by respondents taking the test, these cases reflecting inconsistency can be used as examples (together with the intended code based on coding guidelines) in coder training materials to improve the coding guides and coder training.

In the third step of this analysis, we compared the frequencies of raw responses and unique responses and investigated the pattern of unique responses using visual representations with an eye toward whether human workload could be reduced. If the number of unique response was smaller than the total number of raw responses, this indicated that the workload of human coders could be reduced and the coding process efficiency increased. However, if the number of unique responses approximately equaled the number of raw responses, this indicated that the workload of human coders could not be reduced much. We were able to classify items into groups with different levels of expected gain of the machine-supported coding based on the extent to which regularities or commonalities in the observed responses were observed: large-gain machine-coding with high level of regularities, moderate-gain machine-coding with a medium level of regularities, and small-gain machine-coding with a low level of regularities. In particular, the gain from using machine-supported coding was expected to vary by item and country.

Results

From the 13 items examined, we present the findings of our examination with a focus on the three items shaded in Table 2 regarding the different levels of regularities across the seven countries. To examine the efficiency of the proposed machine-supported coding system, frequencies for each unique response and visual representations are given for each of the three items by country. In accordance with policies regarding confidentiality and item disclosure, only altered responses were used here to illustrate how regularity levels of responses were defined.

High Level of Regularities

Table 3 lists the frequencies of identical response classes for an example item (Item 3) that can be classified as a large-gain machine-coding item with high level of regularities. The three most frequently observed unique responses and their frequencies (class size) are shaded in the table. Table 3 provides frequencies separately by score given: full or no credit. Frequencies of nonresponses are also listed in the rightmost column. For Item 3, there were 1,838 raw responses in the

Table 3 Large-Gain Machine-Coding Item With High Level of Regularities (Country A) (Item 3)

Response	Frequencies (full credit)	Frequencies (no credit)	Frequencies (missing)
30	1,467	0	0
30 minutes	23	0	0
30 mins	7	0	0
…			
10	0	6	0
5	0	6	0
12	*1*[a]	3	0
…			
(No response)	0	0	252
Total	1,509	76	253

Note. Altered responses are shown in this table to illustrate the high level of regularities of responses.
[a]Italics here indicates a miscode.

Country A, and only 50 unique responses were found among them. This implies that it is only necessary for human coders to only code 50 unique responses, or 2.7% (however, potentially they would be scored by multiple coders to ensure agreement) for the same responses to receive the same credit.

For this simple constructed-response item, the answer should be "30" or "30 minutes," and responses including numbers other than 30 should be coded as incorrect. Among all responses, 1,467 students responded correctly with exactly the same response ("30"), and the second most frequently observed unique response was "30 minutes," which came from 23 students. Among responses that received no credit, the most frequently observed were "10" and "5," each of which was observed from six students. Also, we were able to detect a miscode (italicized in the table) from a human coder who gave the wrong score: one student who answered "12" received full credit even though it should have received no credit. This example illustrates how our proposed approach can be utilized to improve the accuracy of coding process by automatically assigning no credit to clearly wrong responses. Finally, 252 students' responses (13.7%) were nonresponses. One incorrect response received a missing code from a human coder although it should have been assigned no credit. These nonresponses represent a large number and could have been coded as no credit without human intervention.

Overall, there were 10 unique responses that were most frequently observed, constituting 81.4% of all 1,838 raw responses. Based on this large proportion of regularities observed in raw responses, we classified this item as large-gain machine-coding item.

Figure 2 illustrates the visual representation of this item across seven country/language groups using bar plots. Each bar plot in Figure 2 shows the cumulative distribution of frequencies of unique responses from each country, with the unique responses sorted by three variables on the x-axis: frequencies of full-credited unique responses on the left, followed by no-credit, and, lastly, nonresponses. The cumulative total of frequencies of these sorted unique responses are shown on the y-axis. Looking at the x-axis, for example in the first plot labeled Country A we see there were 50 unique responses (i.e., equivalent response class) in total out of all raw responses.

At the bottom right corner of the figure, we present a table showing the number of total responses (T) (i.e., the number of respondents) and the number of unique responses (U) found in this certain item. The last row in the table exhibits how much human coding work can be saved (i.e., how much we can gain) if the machine-supported coding system is applied. The percentage of reduction is calculated as $(1 - U/T)$. Note that the additional workload of using multiple human raters (for the purpose of examining coders' reliability) was not taken into account in the calculation. In this example, the workload reduction rate is very high when the machine-supported coding system is used —a range of 94–98% across seven countries.

As listed in Table 3, the most frequently observed response is given in 1,467 full-credited responses. That is indicated on the y-axis of Figure 2 with the first bar reaching 1,467. For this item, there is only a slight increase in the bar heights for the rest of the unique responses. The cumulative total rises only slightly because there are few additional regularities for the rest of the unique responses. In particular, regularities among no-credited responses are very small, making it hard to see the threshold that distinguishes full-credited and no-credited groups. Note that nonresponses constitute one category of the unique responses, with the rightmost bar indicating the nonresponses as listed in Table 3. It can be seen that there is a substantial number of nonresponses, which equals 252 for this country, and is visible with the large jump in cumulative frequencies shown by the rightmost bar. Note that when the sorted unique responses are accumulated, the bar at the rightmost reaches the total number of raw responses, which is 1,838 in this case.

This figure is most helpful in visualizing the pattern of unique responses across countries. Given that different sample sizes are collected in different countries, the number of unique responses and the frequency of occurrence of each unique response will differ systematically by country. As a result, the maximum values on the x-axis and y-axis are not the same by country,[4] though it is clear that the seven countries showed a similar pattern of unique responses (i.e., a high level of regularities in the full-credited unique responses) for this item.

Medium Level of Regularities

Using a structure similar to Table 3, we list the frequencies in Table 4 of homogeneous responses for an example item (Item 2) that can be classified as moderate-gain machine-coding item with medium level of regularities. For this item, we again use Country A as an example. There were 1,815 raw responses in total, with 648 unique responses harvested out of all raw responses. This implies that it is not necessary for human coders to code 1,815 raw responses because of moderate
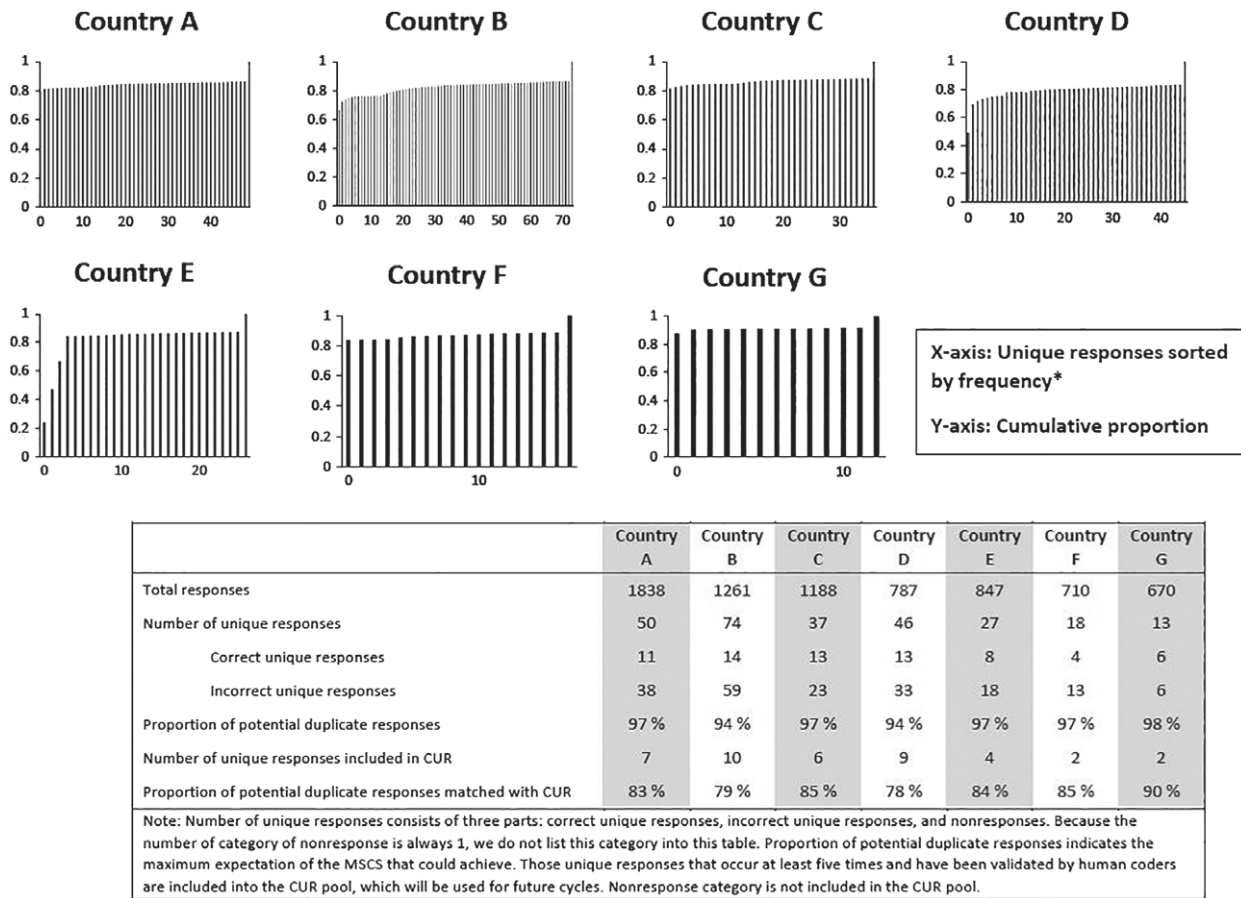
	Country A	Country B	Country C	Country D	Country E	Country F	Country G
Total responses	1838	1261	1188	787	847	710	670
Number of unique responses	50	74	37	46	27	18	13
Correct unique responses	11	14	13	13	8	4	6
Incorrect unique responses	38	59	23	33	18	13	6
Proportion of potential duplicate responses	97 %	94 %	97 %	94 %	97 %	97 %	98 %
Number of unique responses included in CUR	7	10	6	9	4	2	2
Proportion of potential duplicate responses matched with CUR	83 %	79 %	85 %	78 %	84 %	85 %	90 %

Note: Number of unique responses consists of three parts: correct unique responses, incorrect unique responses, and nonresponses. Because the number of category of nonresponse is always 1, we do not list this category into this table. Proportion of potential duplicate responses indicates the maximum expectation of the MSCS that could achieve. Those unique responses that occur at least five times and have been validated by human coders are included into the CUR pool, which will be used for future cycles. Nonresponse category is not included in the CUR pool.

Figure 2 Large-gain machine-coding item with high level of regularities (frequencies sorted left to right by full credit, no credit, and nonresponse). Australia (English), B-S-J-G (China) (Chinese), Germany (German), France (French), Japan (Japanese), Korea (Korean), and the Netherlands (Dutch) are represented by Country A to G in random order.

overlaps; only 648 unique responses, which is 35.7% of raw responses, need to be coded, with all identical responses receiving the same credit.

For this constructed-response item, the answer should be "Earth Road WF" regardless of the capitalization of the letters. Among all responses, 529 students responded correctly with exactly the same response as "Earth Road WF," and the second most frequently observed unique response was "earth road WF" from 76. Moreover, we were able to detect one miscode (italicized in Table 3) from a human coder who gave no credit when the correct answer of "Earth Road WF" was given. This example again illustrates that our proposed approach can be utilized to improve the accuracy of the coding process by automatically assigning full credit to clearly correct responses. Among responses that received no credit, the most frequently observed was "ABC Space Free," seen from 123 students. Unlike the item above that showed small regularities among no-credited unique responses, quite a large number of students provided exactly the same incorrect responses. Finally, 145 students (8%) gave no response. Again, this amount of nonresponses could have been automatically coded as no credit. In addition, the most frequently observed 10 unique responses constituted 49.4% of all 1,815 raw responses. Given that nearly half of responses are covered by these most frequently provided 10 unique responses, we classified this item as moderate-gain machine-coding item.

Figure 3 illustrates the visual representation of this item across seven country/language groups using bar plots and is set up the same as Figure 2. In this case, the first plot, labeled Country A, shows there were 648 unique responses in total out of all raw responses as indicated on the *x*-axis. As listed in Table 4, the most frequently observed response comes from 529 full-credited responses, which is indicated on the *y*-axis with the first bar reaching 529. For this item, there is only a slight increase in the bar heights for the rest of the full-credited unique responses, indicating minimal increase in the cumulative amount. However, a clear jump around 141 on the *x*-axis shows there are regularities among no-credited

Table 4 Moderate-Gain Machine-Coding Item With Medium Level of Regularities (Country A) (Item 2)

Response	Frequencies (full credit)	Frequencies (no credit)	Frequencies (missing)
Earth Road WF	529	*1*[a]	0
earth road WF	76	0	0
earth road wf	45	0	0
…			
ABC Space Free	0	123	0
ABC's Space Free	0	39	0
ABC's space free	0	16	0
…			
(No response)	0	0	145
Total	809	861	145

Note. Altered responses are shown in the table to illustrate the moderate level of regularities of responses.

[a]Italics here indicates a miscode.

Table 5 Small-Gain Machine-Coding Item With Low Level of Regularities (Country A) (Item 11)

Response	Frequencies (full credit)	Frequencies (no credit)	Frequencies (missing)
It states what the paragraph is going to be about.	2	*1*[a]	0
it tells you what the paragraph is about	2	0	0
its telling you what the paragraph is about	2	0	0
…			
don't give up	0	2	0
idk	0	2	0
?	0	1	0
…			
(No response)	0	0	504
Total	1,080	198	504

Note. Altered responses were shown in the table to illustrate the low level of regularities of responses.

[a]Italics here indicates a miscode.

responses as well. This reflects that the most frequently seen no-credit unique response came from 123 students as listed in Table 4. Finally, the jump reflected in the right-hand bar indicates a substantial number of nonresponses, which is 145 students for this country. Note that the bar at the rightmost reaches the total number of raw responses, which is 1,815 in this case. Although the real values on the *x*-axis and bar heights are not the same across countries, it is clear that the seven countries showed a consistent pattern of unique responses (i.e., moderate level of regularities in the full-credited and no-credited unique responses) for this item. The proportion of saved workload is within a range of 39–80% across the countries if the machine-supported coding system were to be applied.

Low Level of Regularities

Following the structure as seen in Tables 3 and 4, Table 5 lists the frequencies of unique responses for the last example item (Item 11), which can be classified as small-gain machine-coding item with low level of regularities. For this item, there were 1,782 raw responses in total from Country A, and 1,274 unique responses were harvested out of all raw responses. Although the number of unique responses seems quite large compared to the two items above, we can still avoid the need to manually score 508 raw responses. Note that among the reduced 508 raw responses, 504 responses (99.2%) are nonresponses as listed in Table 5.

For this constructed-response item, students have to provide the reasonable answer in a sentence, and an insufficient or vague response should be coded as incorrect. Among all responses, the first three full-credited unique responses came from only 2 students, respectively. Regularities in raw responses are rarely observed among no-credited responses. The largest frequencies of unique responses, either in full-credited or no-credited response group, were just 2. However, over a quarter of students, or 504 (28.3% of the total) did not produce a response. Although this item contains only a low level of regularities, a considerable amount of nonresponses itself could have been automatically coded as no credit without
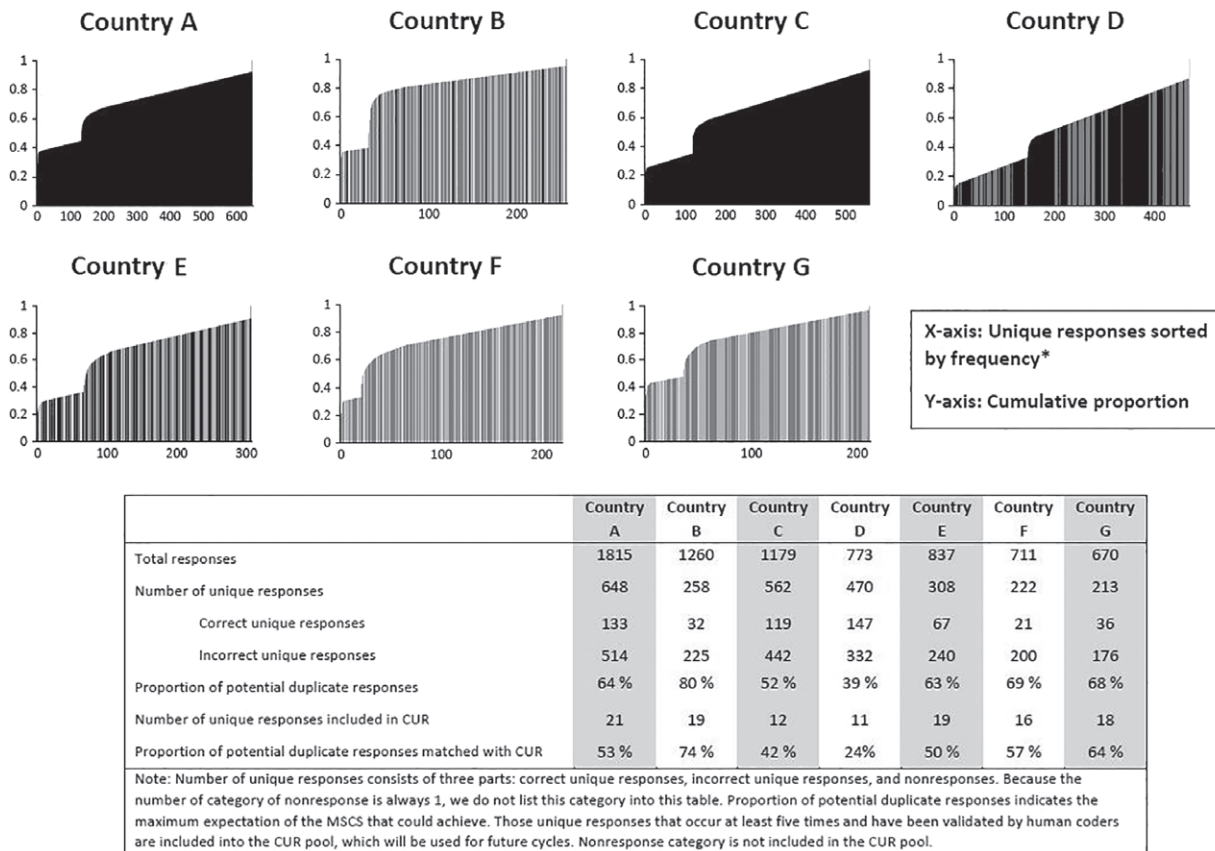
	Country A	Country B	Country C	Country D	Country E	Country F	Country G
Total responses	1815	1260	1179	773	837	711	670
Number of unique responses	648	258	562	470	308	222	213
Correct unique responses	133	32	119	147	67	21	36
Incorrect unique responses	514	225	442	332	240	200	176
Proportion of potential duplicate responses	64 %	80 %	52 %	39 %	63 %	69 %	68 %
Number of unique responses included in CUR	21	19	12	11	19	16	18
Proportion of potential duplicate responses matched with CUR	53 %	74 %	42 %	24%	50 %	57 %	64 %

Note: Number of unique responses consists of three parts: correct unique responses, incorrect unique responses, and nonresponses. Because the number of category of nonresponse is always 1, we do not list this category into this table. Proportion of potential duplicate responses indicates the maximum expectation of the MSCS that could achieve. Those unique responses that occur at least five times and have been validated by human coders are included into the CUR pool, which will be used for future cycles. Nonresponse category is not included in the CUR pool.

Figure 3 Moderate-gain machine-coding item with medium level of regularities (frequencies sorted left to right by full credit, no credit, and nonresponse). The delineation between these three types of responses will be visible only when there is a significant change in frequency. Although this is a bar chart, the outlines of the bars are not visible due to the condensed scale of the horizontal axis, making the bars closely adjacent. Australia (English), B-S-J-G (China) (Chinese), Germany (German), France (French), Japan (Japanese), Korea (Korean), and the Netherlands (Dutch) were represented by Country A to G in random order.

any effort from human coders. Given the low level of regularities observed from raw responses, we classified this item as small-gain machine-coding item.

Figure 4 illustrates the visual representation of this item across seven country/language groups using bar plots and is set up the same as Figures 2 and 3. In this case, the first plot, labeled Country A, shows that there were 1,274 unique responses in total out of all raw responses as indicated on the *x*-axis. As listed in Table 5, the most frequently observed response comes from three full-credited responses, and it is indicated on the *y*-axis that the first bar reaches 2. In particular, for this item, there is no distinctive jump in the bar plot. A straight diagonal line shows until it reaches the rightmost bar representing the nonresponses. Instead, it is seen that there are a large number of nonresponses, which is 504 students for this country, and it is indicated with a jump for the rightmost bar. The rightmost bar reaches up to the total number of raw responses, which is 1,782 in this case, including full-credited responses, no-credited responses, and nonresponses. Although the real values on the *x*-axis and bar heights are not the same across countries, it is clear that seven countries showed a consistent pattern of unique responses (i.e., almost no regularities in the full-credited unique responses and no-credited unique responses) for this item. The proportion of saved workload would be relatively low — a range of 5–29% when the machine-supported coding system is applied.

Note that for items with a low level of regularities, the small gains are mainly contributed by nonresponses rather than identical responses. It implies that how much workload can be reduced in such a small-gain machine-coding item largely depends on the ratio of nonresponses. For instance, there is a high proportion of nonresponse (over 20%) in Country A, as shown in the highest bar to the right end in the Country A plot, and a relatively low missing rate (around 5%) in Country G. That means Country A would have the biggest gain from machine-supported coding system (29%) versus Country G with the smallest (5%).
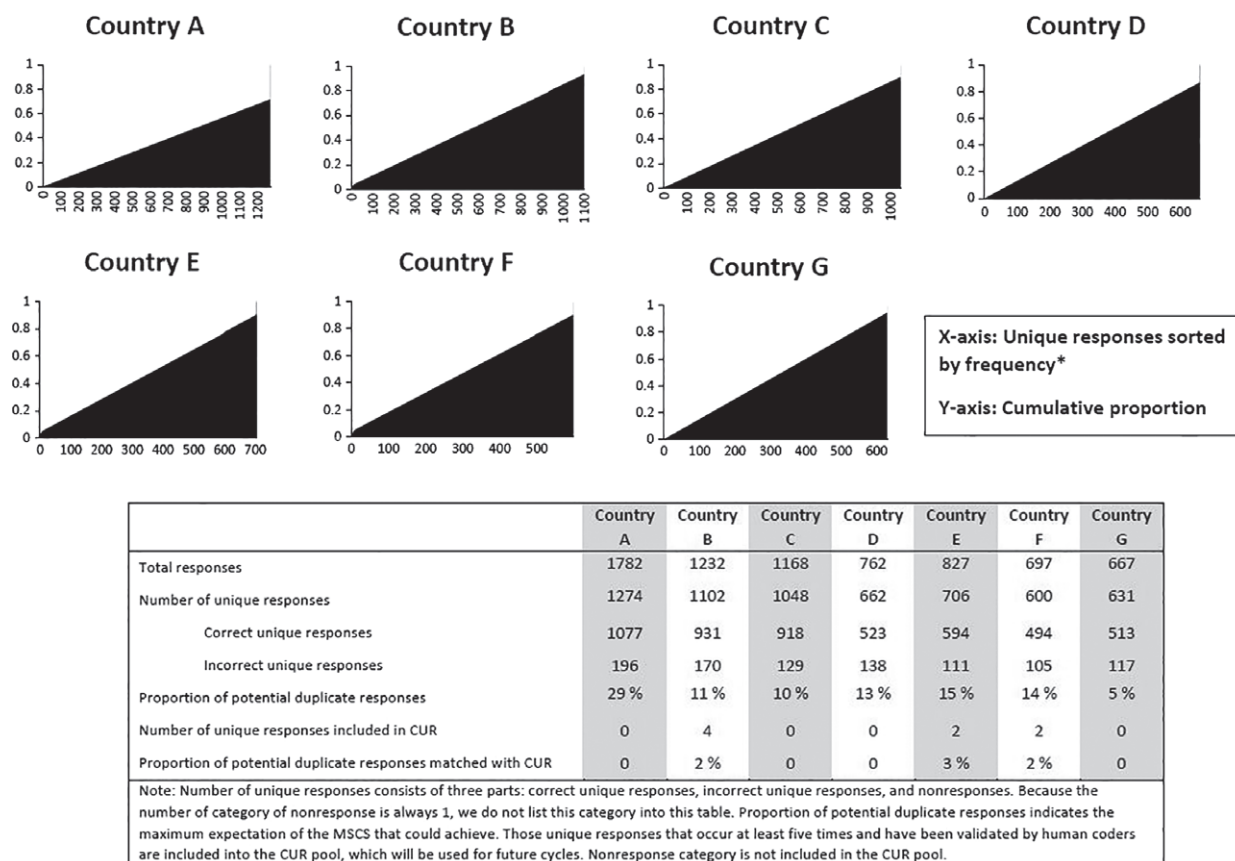
	Country A	Country B	Country C	Country D	Country E	Country F	Country G
Total responses	1782	1232	1168	762	827	697	667
Number of unique responses	1274	1102	1048	662	706	600	631
Correct unique responses	1077	931	918	523	594	494	513
Incorrect unique responses	196	170	129	138	111	105	117
Proportion of potential duplicate responses	29 %	11 %	10 %	13 %	15 %	14 %	5 %
Number of unique responses included in CUR	0	4	0	0	2	2	0
Proportion of potential duplicate responses matched with CUR	0	2 %	0	0	3 %	2 %	0

Note: Number of unique responses consists of three parts: correct unique responses, incorrect unique responses, and nonresponses. Because the number of category of nonresponse is always 1, we do not list this category into this table. Proportion of potential duplicate responses indicates the maximum expectation of the MSCS that could achieve. Those unique responses that occur at least five times and have been validated by human coders are included into the CUR pool, which will be used for future cycles. Nonresponse category is not included in the CUR pool.

Figure 4 Small-gain machine-coding item with low level of regularities (frequencies sorted left to right by full credit, no credit, and nonresponse). The delineation between these three types of responses will be visible only when there is a significant change in frequency. Although this is a bar chart, the outlines of the bars are not visible due to the condensed scale of the horizontal axis, making the bars closely adjacent. Australia (English), B-S-J-G (China) (Chinese), Germany (German), France (French), Japan (Japanese), Korea (Korean), and the Netherlands (Dutch) were represented by Country A to G in random order.

Discussion and Conclusion

In this report, we describe the development of a machine-supported coding approach for constructed-response items in international large-scale assessments such as PISA. There are two major reasons why there is room for improvement in the current human coding process: (a) a lack of consistency among human coder scores, possibly due to lack of understanding of coding rubrics, or coder training, and (b) variation in coding reliability across items and countries. The shift to CBA made it possible to collect all responses using technology and opens avenues to utilize these machine-recorded responses in associated coding procedures, thus, offering the possibility to introduce analysis methods to support coding and improve data processing and analyses in future cycles.

The purpose of our research is to develop a computer-supported coding system to improve the efficiency and accuracy of the coding process for constructed-response items. One important aspect of this approach is generating a pool of unique responses with pre-assigned scores (CUR pool), which helps reducing the human-coding task. This is easily achieved by post-processing the PISA 2015 data in preparation for the 2018 data collection, by extracting unique responses and processing new responses to enhance the existing CUR pool for each item. Because trend items are typically used over three cycles (i.e., one time as the major domain and twice as the minor domain) and PISA implements a field trial before the main survey, the collection of unique responses for the CUR pool is expected to be a powerful tool to considerably reduce the amount of human coding while increasing coding consistency.

In the pilot study, we examined the machine-supported coding system using 13 example items in the reading domain across seven countries with different languages used for testing in PISA 2015. Regarding the accuracy of existing coder data, across seven countries, only a few cases were spotted as miscodes for easy-to-code items (i.e., large-gain items with

high level of regularities), but more miscodes or inconsistent-coding cases were observed for difficult-to-code items (i.e., small-gain items with low level of regularities). This pilot study also provided information on how to revise the coding rubrics and coder training material based on real responses from students. More importantly, by calculating the frequencies of unique responses by full- and no-credit codes, we were able to identify cases where miscodes were assigned or human coders did not agree well. Because all the unique responses are from real responses that students provided during the test, these inconsistently coded cases can be used as examples in coder training materials in order to improve the coding guides and coder training.

In terms of efficiency of the proposed approach, we classified items into three categories: (a) large-gain machine-supported coding with a high level of regularities, (b) moderate-gain machine-supported coding with a medium level of regularities, and (c) small-gain machine-supported coding with a low level of regularities. More specifically, the number of unique responses out of all raw responses became smaller at different magnitudes: As it became more straightforward to do machine-supported coding, fewer unique responses were harvested. It was clearly shown that when high or medium levels of regularities exist among raw responses, machine-supported coding significantly reduces human coders' workload (e.g., more than 90% for the large-gain machine-coding example item). Even when the number of unique responses was similar to the number of raw responses for small-gain machine-supported coding items, the proportion of automatically coded nonresponses helped reduce human coders' workload. This suggests that exclusion of nonresponses could provide time and cost savings for any item. Finally, it is promising that a consistent pattern for each item was observed across the seven countries we examined. This implies the feasibility of our proposed framework and approach in the context of multi-language based international large-scale assessments such as PISA.

While there are apparent benefits from the machine-supported coding system, we also note some limitations. First, a large proportion of missing responses may artificially inflate the gain; a large equivalent response class based on a missing response may give the appearance that the method reduces workload significantly. We chose seven countries with at least average performance in PISA in order to avoid large proportions of nonresponses, but it would be interesting to compare the reduction of workload across countries of different performance levels.

Second, the current study relies on a limited set of countries, observed responses in one data collection, and sample statistics only. The proposed method is a basic approach that can be applied to any language, in that equivalent response classes are based on exact match only. It is a topic for future research to allow for some fuzziness of the response classes (e.g., Sukkarieh, von Davier, & Yamamoto, 2012) or to include preprocessing and base the definition of response classes on strings without white space, punctuation, and capitalization, for example (e.g., Manning & Schütze, 1999).

Third, from the present study, it appears that the items with low-level regularities responses would see very limited reductions of workload from the machine-supported coding system. However, this response group is interesting in its own right, not just to improve the efficiency of scoring. For example, it could be studied whether, after controlling for ability, those regularities are similar across countries as would be expected. Also, it would be interesting to examine whether more substantial workload reduction could be obtained if more advanced machine learning and natural language processing techniques were applied.

In future extensions of this study, we plan to try this approach on more languages used in PISA, such as Arabic, Indian, and Russian, and across all items in all domains, including math, reading, and science. In particular, we will develop indices that capture the expected gain of machine-supported coding in order to: (a) examine whether the pattern of observed unique responses appears consistent across countries, and (b) classify items into different levels of regularities and prioritize the system in operation. In addition, applying simple preprocessing to extract and discard invalid responses (e.g., only punctuations or white spaces in the raw response) could potentially increase the efficiency of the coding process, as could additional clustering. For instance, as with the example shown in Table 4, standardizing the data by removing capitalization would help merge the first three frequently unique responses ("Earth Road WF," "earth road WF," and "earth road wf") into one ("earth road wf") into one equivalent response class and further reduce the workload for coders.

In practice, our target is to implement this proposed machine-supported coding system for the field trial of PISA 2018. The coded responses observed on trend items that have been used and coded in the PISA 2015 main survey will serve as historical data in the CUR pool after a consistency check of the scores in each country. To ensure the reliability of the coded responses in the CUR pool, the frequency of occurrence of one constructed response in an equivalent response class is required to be at least five in the historical dataset. If there is uncertainty on the class label (i.e., disagreement

between human coders), for instance, if the same constructed responses are coded into multiple categories by different human coders, we will suspend such coded responses from the CUR pool.

Following the workflow that was introduced earlier in the Developing a Machine-Supported Coding System in PISA section, the new responses collected in the PISA 2018 field trial will be input into the system and matched with the unique responses stored in the CUR pool. Given an exact matching, the corresponding score stored in the CUR pool will be automatically output; otherwise, the new raw responses will be assigned to (multiple) human coders. The nonresponses will be excluded before being assigned to the human coders at an initial preparation stage. We expect that the machine-supported coding system will significantly reduce the workload for human coders; human coders can then be better assigned to coding tasks that are essential without wasting their time on repetitive work. Further, the potential of more multiple coding for newly observed responses would allow us to examine coder reliability more thoroughly and eventually enhance the accuracy of the coding procedure for the constructed-response items and strengthen the validity and the data quality used in the international large-scale assessments.

In conclusion, our pilot study demonstrates how the proposed machine-supported coding system can be utilized in coding constructed-response items across multiple languages in large-scale assessments. This newly developed coding system holds promise in improving accuracy and efficiency of the coding process for constructed-response items for future cycles and is expected to dramatically reduce the burden for human coders. We believe such a hybrid design in the coding process that combines the strengths of technology with the expertise of human coders can be potentially applied in similar settings and be refined for broader usage in further studies.

Acknowledgments

Notes

1 There are two kinds of coding methods for constructed-response items in PISA, computer- and human-coded. Items with numeric responses (i.e., only numbers, commas, periods, dashes, and backslashes can be entered) and responses involving choices from a drop-down menu or selecting rows of data are coded via computer. All others, typically answered by inputting text-based entries, are coded by human raters.

2 We used two measures—proportion agreement and Cohen's kappa agreement—to examine coder reliability. Proportion agreement is calculated based on the observed count, that is, the total number of responses with agreement between two raters divided by the total number of responses. This measure is easier to understand and, thus, it is easier to communicate the results to participating countries in PISA. Cohen's kappa is more complex but superior statistically because it takes the chance of random agreement into consideration. In calculations on an average level across items and countries, these two measures proved to not differ much.

3 Given the complexity of coding in partial-credited items, we did not include such items in our pilot study.

4 The standardized size of each plot was fixed to illustrate the comparable patterns between the number of unique responses and accumulative frequencies of sorted unique responses across countries. Values vary along the axes as a result.

References

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113–121). Mahwah, NJ: Lawrence Erlbaum Associates.

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities, 37*(4), 389–405. https://doi.org/10.1023/A:1025779619903

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Organisation for Economic Co-operation and Development. (2013). *PISA 2015: Draft reading literacy framework*. Paris, France: Author.

Organisation for Economic Co-operation and Development. (2016). What is PISA? In *PISA 2015 assessment and analytical framework: Science, reading, mathematic and financial literacy*. Paris, France: OECD Publishing.

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

Sukkarieh, J. Z., 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

von Davier, M., Gonzalez, E., Kirsch, I., & Yamamoto, K. (2012). *The role of international large-scale assessments: Perspectives from technology, economy, and educational research.*, New York, NY: Springer.

Erratum

In Research Report no. RR-17-47, some content was included in error. Corrections were made to this report to remove that content. ETS apologizes for this error.

Suggested citation

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

Action Editor: John Mazzeo

Reviewers: Mo Zhang and Isaac Bejar

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