Psychological Test and Assessment Modeling, Volume 60, 2018 (2), 145-164

Development and implementation of a machine-supported coding system for constructed-response items in PISA

Kentaro Yamamoto[1], Qiwei He[1], Hyo Jeong Shin[1] & Matthias von Davier[2]

Abstract

Approximately a third of the Programme for International Student Assessment (PISA) items in the core domains (mathematics, reading, and science) are constructed response and require human coding. This process is time consuming, expensive, and prone to error. The shift in PISA 2015 from paper- to computer-based assessment digitized all responses and associated coding, providing opportunities to introduce technology and analytical methods to improve data processing and analyses in future cycles. The current study explains the framework and approach for improving the accuracy and efficiency of the coding process in constructed-response items for future PISA cycles. Using the frequency distributions, consistencies of responses in coding categories, analysis of coder agreement, and graphic representations, we investigated the efficiency of the proposed machine-supported coding system for all human-coded items across multiple countries using PISA 2015 data and demonstrate how the proposed system was implemented in the PISA 2018 field trial.

Keywords: machine-supported coding, constructed-response items, human coding, large-scale assessments, PISA

[1]*Correspondence concerning this article should be addressed to:* Kentaro Yamamoto, Educational Testing Service, 660 Rosedale Road, 13-E, Princeton, NJ 08541, USA; email: kyamamoto@ets.org

[2]National Board of Medical Examiners

The move toward computer-based assessment (CBA) holds out promise for significant improvements in data quality, leading to greater precision and increased validity (e.g., von Davier, Gonzalez, Kirsch, & Yamamoto, 2012). CBA allows for capturing responses directly into the system for both multiple-choice and constructed-response items. It provides the possibility of automatic scoring for both response types – using scoring keys for multiple choice and machine scoring for constructed responses.

Human coding of constructed responses is time consuming, expensive, and prone to error due to a lack of consistency among human coders. Such coding tasks become burdensome, considering multilingual environments in an international large-scale assessment, such as the Programme for International Student Assessment (PISA). The PISA, given triennially, is one of the largest internationally standardized assessments and is aimed at evaluating education systems worldwide by testing the skills and knowledge of 15-year-old students. In PISA 2018, students representing more than 80 economies in almost 120 languages (including 116 languages in CBA) will participate, with a focus on assessing their capacity to demonstrate preparedness in various domains, particularly reading, mathematics, and science. The core (or major) domain rotates by cycle. In the PISA 2018 cycle, the major domain is reading and will be administered to all students, while the minor domains of science and mathematics will be administered to about a third of the students each. Nearly a third of the items in mathematics and science and about a half in reading domains in PISA 2015 are constructed response and require human coding.[3]

For the first time, PISA 2015 delivered the assessments of all subjects via computer. The shift in PISA 2015 from paper- to CBA digitized all responses and associated coding, providing opportunities to introduce technology and analytical methods to improve data processing and analyses in future cycles.

The current study explains the framework and approach for improving the accuracy and efficiency of the coding process in constructed-response items for future PISA cycles. Specifically, the research questions focus on (1) what is the commonality of correct and incorrect responses by items across country/languages, (2) whether and how much we can take advantages from the computer-supported coding given the small number of unique responses generally found among correct responses, and (3) whether the commonality of responses is consistent across cycles and country/languages. Based on these research findings, we aim at building up a system that could reduce the number of items that have to be coded by human coders. In this paper, we define coding as a process that initially categorizes written responses into discrete classes, thus facilitating scoring in a later step. Using the frequency distributions, consistencies of responses in coding categories, analysis of coder agreement, and graphic representations, we investigated the efficiency of the proposed machine-supported coding system (MSCS) for all human-coded items across multiple countries using PISA 2015 data and demonstrate how the proposed system was implemented in the PISA 2018 field trial. The ability to collect students' raw responses and

[3]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 back slashes 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.

potentially automate the coding of more complex response types – such as extended, constructed answers–is expected to dramatically enhance PISA's overall data quality and has proved effective in its first implementation in the PISA 2018 field trial.

Motivation of developing a machine-supported coding system

Bennett (2011) defined automated scoring as "a large collection of grading approaches that differ dramatically depending upon the constructed-response task being posed and the expected answer." He categorized two general classes of assessment tasks for which automated scoring could be used. The first entails constructed-response tasks that can be graded using exact-matching techniques. For these problems, the scoring challenge is relatively trivial: The correct/incorrect answer(s) are known in advance and can be used to evaluate the quality of the student's response.

The second general class consists of problems for which the responses are too complex to be graded through the exact-matching approach. Automated scoring of complex responses is generally accomplished via a scoring "model." The model extracts features from the student response and uses those features to generate a score, such as the *c-rater*® (Leacock & Chodorow, 2003) and *e-rater*® scoring engines (Burstein, 2003). Tasks may be scored as right or wrong, but in many cases they also can be graded on a partial-credit scale according to a scoring rubric. Such an automated scoring model is typically developed based on one language (e.g., English) to derive accurate scoring in the specific language environment. Because of language diversity in spelling, grammar, wording, and so on, it is very challenging to generalize one single language model to other languages. Given concerns about the multilingual environments in international large-scale assessments, the automated scoring model categorized in the second class by Bennett is less helpful in the current study.

The MSCS typically follows the first class of automated scoring, that is, graded responses with exact-matching techniques based on historical data. The goal of the current system is to avoid repeated coding of the exact same response string by classifying constructed responses into equivalent response classes. For response classes with verified coding, the coding associated with the response class can then be applied to future observations of the identical response, namely, 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, the proposed method relies on human coding and exact matching of previously established classes of responses with newly observed student responses. That means no computer-based classifications or threshold approach are needed; only exactly matching responses receive a coding as previously established based on human coders. Such an exact matching rule could be easily applied to any language in multilingual-based international large-scale assessments such as PISA.

Human coding system in PISA

Due to a lack of consistency among human coders, human coding sometimes results in low coding reliability. In PISA 2015, typically, the number of raw responses to be coded in a single country per language was around 180,000. Assuming 1,000 responses can be coded by a single human coder per day, it would take 180 person days 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. In the PISA 2018 field trial, an average of eight human coders was assigned per country/language in reading for the standard sample size of 1,500 respondents per country. The number of human coders will be increased in the main survey with a bigger student sample size of over 6,000 per country.

Coder reliability in PISA 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 were 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 mathematics, new science items, 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 (Organisation for Economic Co-operation and Development, 2017). For most CBA countries, the standard inter-rater reliability of Cohen's kappa agreement was above 0.9 for all domains (0.97 in mathematics, 0.90 in reading, 0.90 in new science, 0.93 in trend science, and 0.92 in financial literacy).

The following sections describe how the MSCS was developed and implemented as well as its overall performance in the first actual implementation in the PISA 2018 field trial. We first introduce the development of the MSCS, followed by a pilot study to illustrate its function and performance using the responses collected in PISA 2015 (Yamamoto, He, Shin, & von Davier, 2017). Next, the implementation of the MSCS in PISA 2018 field trial is presented with a focus on the development of a coded unique response (CUR) pool. An overview of the performance of the MSCS in PISA 2018 field trial is also reported. Finally, we discuss how to expand the CUR pool and further enhance the reliability and efficiency of the MSCS for future PISA cycles.

Development of a machine-supported coding system

The idea behind the MSCS is to capitalize on the regularity of students' raw responses. Here, "regularity" refers to the extent to which a small number of "unique" responses

represent all students' responses on constructed-response items.[4] For example, high regularity in correct responses means that a relatively small number of unique correct responses represents a large number of correct responses for a given item. In other words, variability among all correct responses for an item is small. In contrast, there can be numerous incorrect responses for a constructed-response item and are easily recognizable–for example, any number other than the correct number. Identical responses (one unique response) should receive the same code when observed a second time, meaning human coding can be replaced by machine coding in such a situation, reducing repetitive coding work performed by humans. Further, machine coding can reduce inaccuracy caused by human coder error (e.g., not understanding the coding rubric, fatigue, not careful enough, etc.) by assigning "verified" codes established from the historic data (i.e., CUR pool). If the verified correct and incorrect codes could be assigned automatically for identical responses, coding the constructed-response items would be much more efficient and accurate as well as less resource intensive for each country.

Raw responses can generally be categorized into two types: (a) responses with verified coding (including nonresponse) and (b) unique responses that require human judgment. In the implementation, response type (a) can be automatically coded based on the CUR pool, while only type (b) needs to be coded by human coders. For instance, if a constructed-response item has 500 identical responses, the human coder should have to code only once for the unique response. The MSCS can code the other 499 instances, resulting in a 99.8 % workload reduction. However, the proportion of workload reduction is item dependent as it depends on the level of response complexity and the consistency of codes given to that unique response. For instance, straightforward responses to short constructed-response items (such as "3 meters" as the response to a question about finding a distance between two points) would more likely result in more consistent codes and, hence, lead to a larger workload reduction than moderately complex responses (such as explanations of how a drug functions).

As Figure 1 shown, the workflow of the MSCS can be divided into two phases: (a) create the CUR pool by identifying the consistently coded frequent responses, and (b) comparing the new responses against the CUR list. In the first phase, historical data – for example, the coded raw responses from the PISA 2015 main survey – are analyzed, and a simple algorithm sorts raw responses by code categories (e.g., 0, 1, 2, 7, and 9). If there is a common code that applies to the sets of identical responses and is exclusive (i.e., if the same response exists in only the "correct" category, but not in the "incorrect" category), a CUR pool can be generated based on the equivalent code and the code is assumed to be verified.

[4]For example, "30m", "30 m", "30 meters" were treated as three "unique" responses, because they are different in terms of spaces or abbreviation used in the raw responses. No preprocessing (e.g., removing spaces) has been conducted for the PISA 2018 field trial.

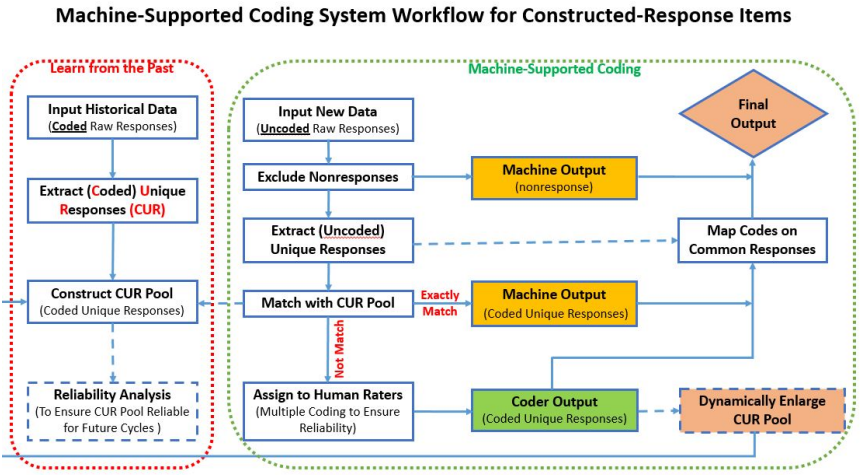


Figure 1:
Machine-supported coding system workflow for constructed-response items

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 in the given country/language group for the PISA 2018 field trial, the stored response code is directly applied to the new respondent's answer. The current MSCS system uses exact response match (including space, spelling mistake, punctuation, etc.) with the CUR pool. Nonresponses such as blanks can be assigned the appropriate nonresponse code. Only those responses that cannot be matched to an identical response stored in the CUR pool are assigned to (multiple) human coders.

Pilot study: Machine-supported coding system in PISA 2015

The potential gain of the MSCS was tested using 13 items from the reading domain in PISA 2015 across seven country/language groups – Australia (English), B-S-J-G (China) (Chinese)[5], France (French), Germany (German), Japan (Japanese), Korea (Korean), and the Netherlands (Dutch) – in a pilot study (Yamamoto et al., 2017). The country/language group set was selected with a diversity in languages and culture: Both alphabetic-based languages (European languages such as English, French, German, and Dutch) and character-based languages (Asian languages such as Chinese, Japanese, and Korean) were represented. In accordance with the policies regarding confidentiality and item disclosure, we anonymized all the countries' names herewith after, instead, used "Country A to G" to

[5]In PISA 2015, only four provinces in China participated the assessment, including Beijing, Shanghai, Jiangsu and Guangdong. We abbreviated this group as "B-S-J-G (China)" to keep consistent with the PISA 2015 technical report.

represent the seven countries in a random order. Also, only "altered" responses were used here to illustrate how regularity levels of responses were defined.

The sample items were selected based on a wide range of regularities of responses. The level of regularities was defined as the ratio between total responses and unique responses per item. Three levels of regularities (i.e., high, medium, and low) were used in the current study. The ratio for an item with a high level of regularity responses was typically more than 20 to 1, meaning one unique response on average represented more than 20 responses in this item. The ratio threshold decreases to 2 to 1 for items with moderate-level regularity responses. When the ratio is lower than 2 to 1, it indicates the item with a low level of regularity responses.

High level of regularities

Table 1 lists the frequencies of identical response classes for an example item that could be classified as a large-gain machine-coding item with high level of regularities. This table provides frequencies separately by score given: full or no credit. Frequencies of nonresponses are also listed in the rightmost column. Using the sample in Country A, there were 1,838 raw responses in this item, with only 50 unique responses were found among them. This implies that human coders would only have been required to code 50 unique responses, or 3 %, for the identical 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 have been coded as incorrect. Among all responses, 1,467 students responded correctly with "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 detected 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 he or she should have received no credit. This example illustrates how our proposed approach can be utilized to improve coding accuracy by automatically assigning no credit to clearly wrong responses. Finally, 252 students' responses (14 %) were nonresponses. One incorrect response received a missing code from a human coder although it should have been assigned no credit.

Table 1:
Large-Gain Machine-Coding Item with High Level of Regularities (Country A) (Item 3)

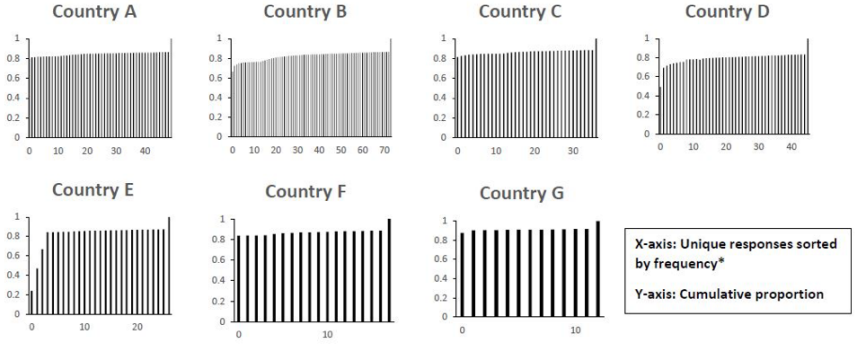
Response	Frequencies (full credit)	Frequencies (no credit)	Frequencies (missing)
30	1,467	0	0
30 minutes	23	0	0
30mins	7	0	0
	…		
10	0	6	0
5	0	6	0
12	*1*[1]	3	0
	…		
(No response)	0	0	252
Total	1,509	76	253

Note.[1] Italics here indicates a miscode. "Altered" responses were shown in the table to illustrate the high level of regularities of responses.

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 proportion of unique responses by each country (vertical axis), with the frequency of unique responses sorted by three categories on the horizontal axis from left to right: full-credited unique responses, no credit, and nonresponses.

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) (i.e., the sum of unique responses in correct and incorrect groups) in the items. The following row "proportion of potential duplicate responses" exhibits the maximum expectation (i.e., upper boundary) that the duplicate responses can be removed from human coding workload if a machine coding engine is applied. The percentage of reduction is calculated as $\frac{1-U}{T}$. Note that the additional workload of using multiple human raters (for examining coders' reliability) was not considered in the calculation. In this high-level-regularity-response example, the proportion of potential duplicate responses is very high if the MSCS is used – a range of 94-98 % across seven countries. The last two rows present the number of unique responses that satisfy the rules to be included in the CUR pool, and the proportion of potential duplicate responses that can be matched in the CUR, which could be regarded as a lower boundary as the minimum expectation from the MSCS. The major rules applied to building up the CUR pool will be addressed in more details in the next section. The nonresponse and unique responses with frequency not less than five times in one and only one coding category were included in the CUR pool. It is noticeable that in such a high-level-regularity-response item, the CUR unique responses are very powerful to save a large proportion of duplicate coding tasks from human coders. Especially in Country G, 90 % duplicate coding tasks could be saved by only two unique responses.

As listed in Table 1, for this item in the Country A, the most frequently observed response was given in 1,467 full-credited responses (80 %). That is the starting point in Figure 2 from the first bar. It is notable that the cumulative proportion rises slowly after the first bar, implying there are few additional regularities for the rest of the unique responses. 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 1. There are a substantial number of nonresponses, which is 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. The cumulative distributions follow similar patterns across countries, meaning there is not a large language effect in this item. The efficiency benefits from the MSCS are consistent across countries for this item.



	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 example item with high level of regularities. *Frequency of unique responses on horizontal axis 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) were represented by Country A to G in a random order.

Medium level of regularities

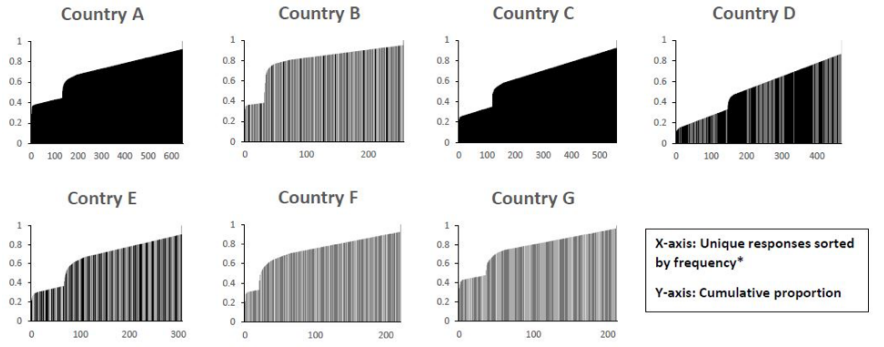
Following a similar structure, we presented an example item with a medium level of regularities response in Table 2. In Country A, there were 1,815 raw responses in total, with 648 unique responses harvested, suggesting only 36 % responses needed to be coded by human coders.

The correct answer for this particular item should be "Earth Road WF" regardless of the capitalization of the letters. Among all responses, 529 students responded correctly with the same exact response as "Earth Road WF," and the second-most frequently observed unique response was "earth road WF" from 76. Moreover, we detected one miscode (italicized in Table 3) from a human coder who gave no credit when the correct answer of "Earth Road WF" was given. Unlike the item above that showed small regularities among no-credited unique responses, many students provided exactly the same incorrect responses.

Table 2:
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*[1]	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. [1]Italics here indicates a miscode. "Altered" responses were shown in the table to illustrate the high level of regularities of responses.



	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. *Frequency of unique responses on horizontal axis 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), the Netherlands (Dutch) were represented by Country A to G in a random order.

Consistent with Figure 2, similar patterns of unique response distributions were found in the item with medium level of regularity responses as illustrated in Figure 3. The most frequently observed response came from 529 (29 %) full-credited responses, which is the starting point in the cumulative proportion axis. The bar heights are slightly increasing for the rest of the full-credited unique responses but followed by a clear jump when the no-credit unique responses joined. The final jump reflected in the right-hand bar indicates a substantial number of nonresponses. The proportion of items not needing human coding is within a range of 39-80 % across countries if the MSCS were to be applied. It was also interesting to find that compared with the high-level-regularity-response item in Figure 2, the number of unique responses included in the CUR was increased in this medium-level-regularity-response item. However, the proportion of potential duplicate responses that matched with the CUR pool was lower across all the country/language groups, meaning the CUR unique responses are a bit weaker compared with the previous example item on account of a relatively lower frequency of each CUR unique response.

Low level of regularities

Following the same structure, Table 3 lists the frequencies of unique responses for the last example item that can be classified as a 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 could still avoid the need to manually score 508 raw responses. Note that among the reduced 508 raw responses, 504 responses (99.2 %) were nonresponses, as listed in Table 3.

For this constructed-response item, students needed to provide a reasonable answer in a sentence; an insufficient or vague response should have been coded as incorrect. Among all responses, the first three full-credited unique responses came from only two students, respectively. Regularities in raw responses were rarely observed among no-credited responses. The largest frequencies of unique responses, either in the full-credited or no-credited response group, were just two. However, over a quarter of students, or 504 (28.3 % of the total), did not produce a response. Although this item contained only a low level of regularities, a considerable number of nonresponses could have been automatically coded.

Analogous to illustrations in figures 2 and 3, we found similar patterns of cumulative distribution in the example item with low level of regularity responses in Figure 4. In Country A, the most frequently observed response came from three full-credited responses. A straight diagonal line is observed until it reaches the rightmost bar, suggesting almost all responses were unique. A high jump in the rightmost bar is spotted for a high nonresponse rate in this item. The proportion of saved workload would be relatively low – a range of 5-29 % if the MSCS were applied.

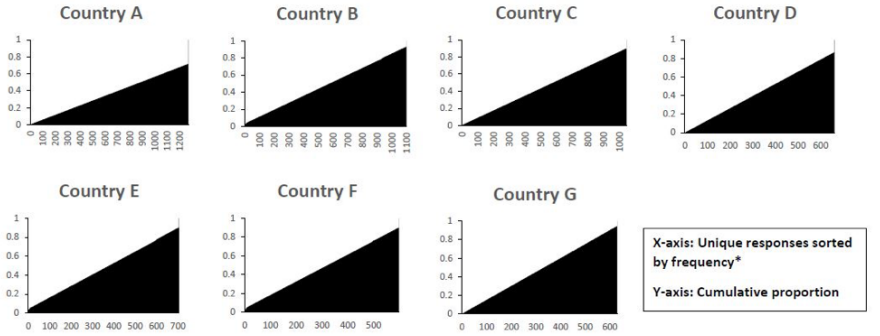
Table 3:
Small-Gain Machine-Coding Item with Low Level of Regularities (Country A) (Item 11)

Response	Frequencies (full credit)	Frequencies (no credit)	Frequencies (missing)
It states what the paper is going to be about.	2	*1*[1]	0
it tells you what the paper is about	2	0	0
its telling you what the paper is about	2	0	0
…			
don give up	0	2	0
Idk	0	2	0
?	0	1	0
…			
(No response)	0	0	504
Total	1080	198	504

Note. [1]Italics here indicates a miscode. "Altered" responses were shown in the table to illustrate the high level of regularities of responses.

For items with a low level of regularities, the small gains are mainly contributed by non-responses rather than identical raw responses, implying that the potential decrease in workload from a small-gain machine-coding item largely depends on the ratio of nonresponses. For instance, there was 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, while there was a relatively low missing rate (around 5 %) for Country G, suggesting Country A would benefit more from the MSCS than the Country G merely by nonresponse rate. We also noticed that the unique responses that could be included into the CUR pool became rare in the low-level-regularity-response item. Due to the extremely low frequency of each unique response, the removal of duplicate responses could not be benefited much from the CUR pool.

To sum up, in this pilot study, the sample item with the most instances of repeated raw responses resulted in a maximum expectation of 94-98 % workload reduction across country/language groups, whereas the sample item with the fewest repeated responses reduced coding workload by as little as 5-29 %. More importantly, when items were categorized into three groups in terms of regularities – high, medium, and low – there was a fairly consistent pattern in item categorization across many country/language groups. These results indicate that it is feasible to increase the usage of MSCS for PISA, which has more than 80 countries and 100 language versions. The results from the pilot study also suggest that it is possible to use the MSCS for the completely new constructed-response items (without any historical data) by having empty responses as one unique response. More specifically, an algorithm can evaluate whether a new response was observed in the CUR, even if the CUR is initially a nonresponse. Any new, unique response not in the CUR will be a new one and be presented to a human coder. If multiple coders all agree in terms of the assigned response (typically more than two) for any such response, it is possible to add the verified unique response and its associated code to the CUR for the future cycle as a standard step.



	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. *Frequency of unique responses on horizontal axis 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), the Netherlands (Dutch) were represented by Country A to G in a random order.

Implementation of machine-supported coding system in PISA 2018 field trial

Preparation of MSCS for PISA 2018 field trial

In preparation for the PISA 2018 field trial, the MSCS was applied to all constructed-response items across all domains based on extracted data from the PISA 2015 main survey. It showed that across all items and country/language groups, the percentages of identical responses among all responses constituted approximately 40 % in mathematics, 28 % in reading, 22 % in science, and 18 % in financial literacy, meaning the human workload could potentially be reduced by those amounts. Raw responses from a total of 146 items (21 items from math, 58 from science, 51 from reading, and 16 from financial literacy) across 59 countries were used to prepare the PISA 2018 field trial CUR pool. In the CUR pool, each unique response was associated with a verified code (i.e., 0, 1, 2, 9, etc.), which

was consistent with the coding guidelines from PISA 2015. The CUR pool was built to be country/language-specific; within the CUR pool, coded unique responses were stored separately by domains and language groups.

Two major rules were used when the unique responses were extracted and entered into the CUR pool. First, the response to an item in a specific country/language group should occur at least five times in one coding category. To ensure that the CUR pool contained accurate and verified codes for each unique response, only unique responses with identical and exclusive codes were included. The second rule was set for the nonresponse category. An empty response was added to each item regardless of the frequency of nonresponses. This approach ensured at least one unique raw response (i.e., empty response) could be found in each constructed-response item in the CUR pool, meaning the nonresponse could be directly filtered and coded by the machine rather than assigned to human coders.

The PISA 2015 data was used to build and verify all the coded unique responses within the current CUR pool. The new raw responses collected in the PISA 2018 field trial were added into the MSCS and compared with the verified CUR on an item-by-item basis for each country/language group. Once a new response was found with an exact match to an identical CUR to a specific item, the stored code in the CUR pool was automatically applied to this response. During the PISA 2018 field trial, the responses that could not be matched with the existing CUR pool as well as the responses collected for the new items were assigned to human coders. These items will be examined after the field trial to decide whether they can be added into the CUR pool. By repeating this process, the CUR pool can be expanded, further verified, and prepared for the PISA 2018 main survey and future cycles.

Performance of MSCS in PISA 2018 field trial

The PISA 2018 field trial used the newly developed MSCS, based on PISA 2015 data, for the first time as part of the coding process for the constructed-response items. The system was applied for all country/languages groups that participated in the PISA 2018 field trial, except for some country/language groups that are either new to PISA or switching from paper- to computer-based assessment. Due to having no historical data in the CUR pool, they were not eligible for this system.

The performance of the MSCS was evaluated with respect to the efficiency of the system and its capability to monitor and improve coding accuracy. As to efficiency, various types of automatically coded responses were summarized across items and country/language groups. Before the system existed, of course, all student responses, including empty responses, were assigned to human coders without exception. Thus, this evaluation revealed the extent to which the burden of human coding in the CBA was decreased in the 2018 field trial. As for the capability to monitor the accuracy of human coding, consistency of human-coded responses was examined relating codes to students' raw responses.

Table 4:
Efficiency of Machine-Supported Coding System Implemented in Constructed-Response Items in PISA 2018 Field Trial

	Machine-coded			Human-coded
	Total	Missing	Valid	
Mathematics	34 %	17 %	17 %	66 %
New Reading	16 %	11 %	5 %	84 %
Trend Reading	21 %	10 %	11 %	79 %
Science	25 %	13 %	12 %	75 %
Financial Literacy	13 %	12 %	1 %	87 %

Table 4 summarizes the average efficiency of the MSCS across all items and country/language groups in each domain. On average, the proportion of items not needing human coding was reduced for the 2018 field trial from a low of approximately 13 % in financial literacy to a high of 34 % in mathematics. To clarify the efficiency given by different sources, we calculated the empty (missing) responses and valid responses separately. On average, approximately 10-17 % of the total responses in trend reading and mathematics, respectively, were empty responses and automatically coded by the system. The MSCS was also efficient for new items in reading, where no historic data were available, reducing the proportion of items that used to be coded by human coders by 11 % on average just by excluding blank responses. For the valid responses, approximately 0.8 % in financial literacy to 17 % in mathematics efficiency was gained. For new reading items, the proportion of items that used to be coded by human coders was reduced by an additional 5 % on average by incorporating obviously incorrect responses for some item types (e.g., responses where a student selected a radio button option but typed no text in the text box were coded as "incorrect"). The proportions of the human-coded responses are shown in the last column. These values correspond to the proportions of responses where the current system could not find an exact match to the raw responses in the current CUR pool that was built based on the 2015 main survey. New items (especially in the reading and financial literacy domains in the 2018 field trial) and new countries that were not included in the PISA 2015 do not have a CUR pool due to the absence of historical data, so no efficiency could be gained. It is also the main reason that a gap was observed between the theoretical maximum gains in efficiency expected based on PISA 2015 and actual implementation in the 2018 field trial. On average, approximately 66 % for mathematics to 87 % for financial literacy of the responses had to be scored by human coders in the 2018 field trial after the MSCS was implemented.

As the cycle of assessments proceeds, the CUR pool is expected to grow and the proportion needing human coding is expected to decrease as responses from the 2018 field trial data are added to the existing CUR pool from 2015 main survey data. Furthermore, considering the major domain in PISA 2015 was science while in PISA 2018 it will be a different domain (reading), more constructed-response items are expected to be used in PISA 2018, which would enhance the harvest of the CUR pool even further.

Accurate and reliable coding of item responses, especially for human-coded constructed-response items, is a key component of quality control and is a necessary step for ensuring

valid and comparable assessment results. Before the introduction of CBA, monitoring the accuracy of human-coded responses was resource intensive. CBA enables the capture of students' raw responses and associating these responses to the corresponding codes given by human coders as well as the CUR pool for machine coding. Because the MSCS system decreased the number of human-coded responses by excluding empty responses and machine coding others, more responses could be assigned to multiple coding in the 2018 field trial. This allowed for better monitoring of coding accuracy, not only by comparing results from multiple human coders but by evaluating the assignment of codes to students' raw responses. In addition, increasing the number of verified codes for more complex responses that were validated through multiple coding to the CUR pool further improved the validity of the codes.

Discussion and Conclusion

This paper describes the development and implementation of a machine-supported coding system for constructed-response items in multilingual-based 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 opened 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 need for human coding. 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 part of the major domain and twice as part of 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.

To illustrate the function and performance of the MSCS, we conducted a pilot study in which the MSCS was examined by 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, but more miscodes or inconsistent-coding cases were observed for difficult-to-code items.

In terms of efficiency of the proposed MSCS, 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 reduced 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 can 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.

In addition, our research 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 could identify cases where miscodes were assigned or human coders did not agree sufficiently. 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 to improve the coding guides and training.

As expected, the application of the newly developed MSCS to the PISA 2018 field trial significantly reduced the proportion of items that used to be coded by human coders: from a low of approximately 13 % in financial literacy to a high of about 34 % in mathematics. Thus, both accuracy as well as efficiency of coding was improved. In addition, the system has the capacity to monitor coding accuracy by comparing codes from multiple human coders and assigning these given codes to new students' raw responses.

While there are apparent benefits from the MSCS, we also note some limitations. First, the current CUR pool (for the PISA 2018 field trial) has been established based on a data-driven consistency notion that coherent codes assigned to frequently observed responses would be accurate. However, there is a challenge in validating the accuracy of codes, particularly when the unique response is confusing and difficult to agree upon. This means that unique responses, especially those that were flagged due to low reliability across coders, are recommended to be coded and validated by master coders by country/language groups before being added to the CUR pool. It would be of importance in expanding the CUR pool to improve the efficiency of the MSCS for the future.

Secondly, the current MSCS is built upon specific country/language groups, meaning the languages are not clustered across countries (i.e., the Canada/English group is treated separately from US/English even though the same language is used). It would be more efficient to combine the CUR pool by language groups to further enhance the harvest of unique responses in the language cluster. Further, the proposed MSCS is a basic approach that can be applied to any language, in which 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 (e.g., Manning & Schütze, 1999).

Thirdly, from the present study, it appears that items with low-level regularities responses would see very limited reductions of workload from the MSCS. However, this response group is still of interest, and 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 one might expect. 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.

Finally, the current MSCS assigns human coders only if the new responses were not scored by machine. Hence, direct comparisons between the machine and humans were not available. To monitor the accuracy of the CUR pool, enabling direct comparison between machine and human coders can be considered, for instance, in the PISA 2018 main survey.

In conclusion, along with the pilot study and results reported in research report based on the PISA 2015 main survey, application of the system to PISA 2018 field trial proves the feasibility of the proposed MSCS and provides evidence for improving accuracy and efficiency of the coding process for constructed-response items. Hence, the implementation of this system is recommended for the PISA 2018 main survey and beyond. Also, the MSCS is designed not only for multilingual tests but can easily be adapted to single-language tests as well, reducing redundancy wherever duplicate constructed responses are observed. Moreover, the CUR pool does not have to be static, it can be adaptive within a duration of coding responses. A CUR pool cumulated from previous response data can be dynamically updated when the frequency of new unique responses with consistent coding reaches a certain statistical threshold. Therefore, we believe the MSCS holds promise in a broad range of applications for automatic coding of constructed responses.

References

Bennett, R. E. (2011). Automated scoring of constructed-response literacy and mathematics items. *Advancing Consortium Assessment Reform (ACAR)*. Washington, DC: Arabella Philanthropic Advisors.

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*, 389-405. doi: 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 (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.* Paris, France: OECD Publishing.

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. doi: 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.

Yamamoto, K., He, Q., Shin, H. J., & 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. doi: 10.1002/ets2.12169