Psychological Test and Assessment Modeling, Volume 60, 2018 (2), 141-144

# *Guest Editorial*
# Special Topic: Advances in Educational Measurement

*Andreas Frey[1,2], Christoph König[1] & Christian Spoden[1]*

The term *Educational Measurement* refers to the process of representing differences between persons or other entities in educational contexts in terms of numbers. This includes theory, research, and application concerning study designs, instruments, data collection, statistical analysis, and the usage of the results obtained. Educational measurement has a substantial overlap with psychometrics. The main distinction between educational measurement and psychometrics lies in the content typically focused on; educational measurement is concerned with educational aspects and psychometrics with internal psychological processes (see Jones & Thissen, 2007, for a historical overview of psychometrics).

These educational aspects are currently undergoing major transformations due to the growing importance of digital devices. Today, digital devices already have a strong influence on educational processes and it is likely that this will continue in the years to come. It is not a very daring prediction to state that the next few decades will bring substantial changes to the conditions under which we learn, to what we learn, how we learn, and how we use what we have learned. A general trend in the changes taking place is already evident: Educational processes are becoming more personalized, more flexible, and less standardized. This trend has the potential to promote the quality of education, but it also poses a major challenge to educational measurement. This is the case because standardization and the use of structured processes are key elements of educational measurement. Thus, the compatibility between traditional methods of educational measurement and the current transformations in education is limited. Nevertheless, educational measurement is actively taking up the challenges connected with the growing importance of digital devices in education by conducting research on more personalized and flexible methods of measurement, on statistical modeling, and on using the generated results.

The current special topic of *Psychological Test and Assessment Modeling* presents a series of such research studies. The papers can be grouped into four categories:

[1]*Correspondence concerning this article should be addressed to:* Andreas Frey, Institute of Educational Science, Department of Research Methods in Education, Friedrich Schiller University Jena, Am Planetarium 4, 07743 Jena, Germany. email: andreas.frey@uni-jena.de
[2]Centre for Educational Measurement (CEMO) at the University of Oslo, Norway

(1) Computerized Adaptive Testing and Multistage Testing
(2) Analysis of Large-Scale Assessment Data
(3) Multilevel Structural Equation Modeling
(4) Bayesian Modeling

The first category is devoted to the measurement process itself, the second to a special kind of assessment that aims to draw inferences at the population level, and the last two to complex statistical modeling approaches.

The first two papers of the present issue focus on recent advancements regarding the *analysis of large-scale assessment data*. In the first paper, Yamamoto, He, Shin, and von Davier describe a novel machine-supported coding system for answers given to constructed-response items. The new coding system was developed for the Programme for International Student Assessment (PISA), which switched from paper- to computer-based assessment in 2015. The paper presents brand new information and results from the application of the method in the PISA 2018 field trial. The results underline the feasibility of the proposed machine-supported coding system and provide evidence for its capacity to significantly improve the accuracy and efficiency of the coding process for constructed-response items.

The second paper by Nagy, Nagengast, Becker, Rose, and Frey focuses on the topical issue of item position effects. Item position effects are variations in item parameter estimates with respect to the position in which items are presented to test takers. A common finding across content areas and age groups is that performance items tend to become more difficult towards the end of tests. It is–however–not yet clear which variables stand behind item position effects. The paper of Nagy et al. (2018) is the first publication to analyze such individual correlates of item position effects in a reading comprehension test. The authors propose an item response theory (IRT) model with random effects for the item difficulties and fixed effects for the item discriminations, and they provide an M*plus* syntax for its estimation. As expected, item position effects regarding item difficulties and item discriminations were found. The effects on the item difficulties were systematically related to students' decoding speed and reading enjoyment. Expanding the literature, they analyze and discuss how inferences drawn from test scores are affected by item position effects.

The next two contributions are devoted to recent developments in the area of *multilevel structural equation modeling*. The paper by Kiefer, Rosseel, Wiese, and Mayer proposes a multilevel latent growth components model to account for potentially nonlinear shapes of educational trajectories, a multilevel data structure, and the measurement of unobservable latent constructs. In an empirical illustration, the model is applied to predict the nonlinear development of students' satisfaction with their academic success, based on data from the National Educational Panel Study (NEPS) in Germany. The results indicate that the latent satisfaction of students increased after the first wave and after that remained relatively constant on average, although variation existed both across study programs and individuals. This variation was predicted by the change of major after the first year and by the examination burden. The authors provide a lavaan and an M*plus* syntax so that interested readers can directly estimate their model.

In the fourth paper, Spoden and Fricke investigate the dimensional structure of the classroom management skills of physics and science teachers. Classroom management skills are an important aspect of instructional quality and a key competence of a teacher. Due to the multilevel nature of classroom management skills, however, considerable challenges have to be overcome with regard to the interpretation of the constructs under investigation. Taking up these challenges, Spoden and Fricke apply a shared cluster construct approach to measuring classroom management skills. They identify a three-dimensional structure of classroom management skills, in contrast to the unitary definitions of this construct in other recent studies. The shared cluster construct approach applied in their study illustrates how complex multidimensional indicators of instructional quality can be measured in a psychometrically sound manner in multilevel contexts.

The last two contributions of the first part of the special topic deal with the possibilities of *Bayesian modeling* in educational contexts. In the fifth paper, Trendtel and Robitzsch analyze linear and nonlinear patterns of item position effects, the stability of the effects across different test cycles, and whether item position effects are affected by changes in the test administration mode from paper-pencil testing to computer-based testing. For this purpose, the authors propose a Bayesian IRT model, which is also extended to weighted clustered samples. They applied the model to study item position effects in reading data from PISA 2009, 2012, and 2015. The results from the six countries analyzed provide evidence for linear and nonlinear patterns, stable and instable item position effects, as well as a decrease in the effects caused by a change in the test administration mode in most but not all countries.

The sixth paper by Helm addresses the potential benefits of Bayesian modeling for multilevel latent contextual models in small samples. More specifically, the study focuses on doubly latent multilevel models as state-of-the-art representations of instructional quality. Given their doubly latent specification, these kinds of models pose considerable challenges in terms of sample size. Bayesian modeling offers an approach to meet these challenges; its full potential, however, can only be utilized if background knowledge is introduced into the analysis in the form of informative prior distributions. Accordingly, Helm presents a comprehensive simulation study in which he compares the performance of Maximum Likelihood and Bayesian estimation of doubly latent multilevel models in terms of the accuracy of the group-level effect. In line with previous research, he shows that accurate estimates of the group-level effect are obtained even in the smallest sample sizes when Bayesian estimation with either weakly or fully informative prior distributions is used. An illustration of how to use data from the large-scale assessment Trends in International Mathematics and Science Study (TIMSS) to obtain informative prior distributions makes this paper an important contribution to applied Bayesian modeling in the field of educational measurement.

The six papers assembled in this issue are the first part of the special topic. The special topic will be completed by three more articles that will appear in the next issue of *Psychological Test and Assessment Modeling*.

## References

Helm, C. (2018). How many classes are needed to assess effects of instructional quality? A Monte Carlo simulation of the performance of frequentist and Bayesian multilevel latent contextual models. *Psychological Test and Assessment Modeling*, *60(2)*, 265–285.

Jones, L. V., & Thissen, D. A. (2007). A history and overview of psychometrics. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, volume 26: Psychometrics* (pp. 1–27). New York, NY: Elsevier.

Kiefer, C., Rosseel, Y., Wiese, B., & Mayer, A. (2018). Modeling and predicting non-linear changes in educational trajectories. The multilevel latent growth components approach. *Psychological Test and Assessment Modeling*, *60(2)*, 189–221.

Nagy, G., Rose, N., Frey, A., Becker, M., & Nagengast, B. (2018). Item position effects in a reading comprehension test: An IRT study of individual differences and individual correlates. *Psychological Test and Assessment Modeling*, *60(2)*, 165–187.

Spoden, C., & Fricke, K. (2018). Measurement of teachers' reactive, preventive and proactive classroom management skills by student ratings – Results from a two-level confirmatory factor analysis. *Psychological Test and Assessment Modeling*, *60(2)*, 223–240.

Trendtel, M., & Robitzsch, A. (2018). Modeling item position effects with a Bayesian item response model applied to PISA 2009–2015 data. *Psychological Test and Assessment Modeling*, *60(2)*, 241–263.

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

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.

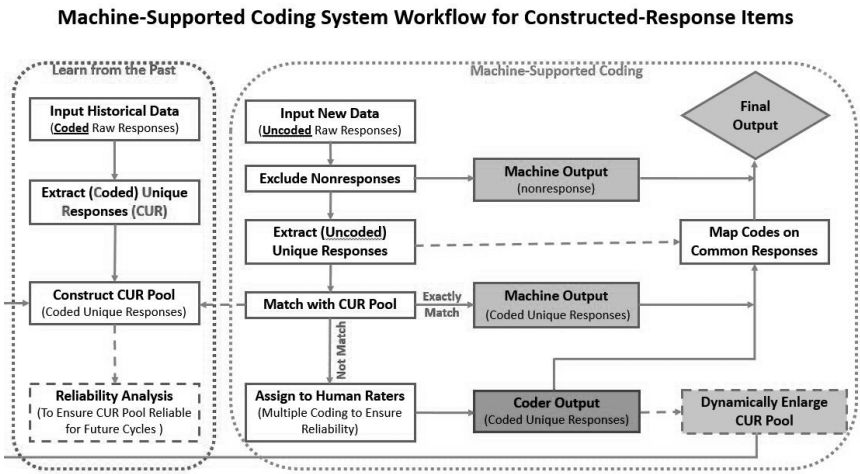


**Figure 1:**
Machine-supported coding system workflow for constructed-response items

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.