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.