Guest editorial

Foreword

The papers in this special issue are based on presentations at a two-day international seminar on managing the quality of data collection in large-scale assessments. The seminar was held on May 11-12, 2017, at the Organisation for Economic Co-operation and Development (OECD) headquarters in Paris. The purpose of this event was to bring together psychometricians and survey methodologists to discuss issues around the identification, treatment and prevention of errors associated with data collections in large-scale assessments, as well as prospects for the evolution of data collection methods. This was the first of a planned series of seminars on methodological issues relevant to the Programme for the International Assessment of Adult Competencies (PIAAC) and other international large-scale assessments such as the OECD's Programme for International Student Assessment (PISA).

The topic of managing quality in data collection was chosen for two main reasons: first, data collection and field operations represent major sources of potential error in any large-scale survey, particularly those such as PIAAC that administer questionnaires and tests of cognitive skills (literacy, numeracy and problem-solving) using interviewer-based methods. The behavior and skills of interviewers, the setting and conditions in which the interview or assessment takes place, the dispositions and motivation of respondents and the capability and capacity of survey organizations collecting the data all have effects on data quality. These effects range from data falsification at the different levels of survey operations at one extreme to satisficing (Krosnick, 1991) by both interviewers and respondents at the other.

Second, developments in information and communications technologies offer an opportunity to achieve considerable improvements in data quality through the identification, treatment and prevention of errors. New technologies are also opening up potential new avenues for data collection. The use of computer-aided personal interviewing and computer-based testing has already led to demonstrable improvements in data quality in large-scale survey assessments and other testing programs. Automatic scoring and automatic range checks for responses reduce the chance for human error, for example. The availability of process data that represent the interactions between interviewers and respondents and the interview/testing application, including timestamps, provides a rich source of information for detecting problems and potentially adjusting for them.

This seminar on managing the quality of data collection in international large-scale assessments was organized into five sessions, each including two presentations. The sessions were preceded by an introduction and overview presentation focusing on the variability and potential sources of errors in comparative surveys. This overview discussed the concept of total survey error (TSE; Groves and Lyberg, 2010), first introduced by Hansen *et al.* (1951, 1953), as a framework for identifying and addressing various sources of non-sampling variance.

The first session included two presentations dealing with understanding the interview process as a way of orienting researchers to some of the special circumstances associated with delivering surveys through face-to-face interviews. This was followed by two presentations with a focus on data fabrication. A third session focusing on the topic of survey error completed the first day of the seminar.

The second day opened with a session on detecting errors during the interview. The first paper focused on the development and use of dashboards to provide key indicators during the data collection process, whereas the second dealt more specifically with interviewer quality control and quality assurance issues. The final session looked toward the future of data collection. It opened with a presentation focused on whether face-to-face interviewing would continue to remain the gold standard for large-scale surveys, whereas the second and final presentation explored new strategies that can either augment or replace traditional approaches to data collection.

An important challenge for those who design and manage international large-scale assessments is to apply what has been learned so far to the design of new tools and systems to facilitate a more preventative and anticipatory approach to quality control and quality assurance in data collection. In the longer term, the availability of new tools and approaches may have profound impacts on how sample surveys and large-scale assessments are conducted. The papers discussed during the event and contained in this special issue represent an important contribution to current thinking about these issues. As one of the presenters noted, the design and implementation of the first cycle of PIAAC in 2012 reflected best practice at the time. However, as we move toward the second cycle, and as the field of international large-scale assessments has advanced in significant ways over recent years, possible improvements to the application of TSE principles in future large-scale assessments will need to be addressed.

Some readers may ask why managing the quality of data collection in large-scale assessments has been chosen as a topic for the journal *Quality Assurance in Education*. The primary reason is that large-scale assessments – both national and international – are an increasingly important source of information for understanding the outcomes of education and training systems. Educators, educational researchers, economists and policymakers look to surveys such as PIAAC to gather important information about the relationship of skill development and life outcomes of individuals, as well as the success of economies. The growing participation in OECD studies such as PIAAC and PISA provides an eloquent testimony to the value of these data. Moreover, the architecture proposed for monitoring progress toward the United Nations Sustainable Development Goals for Education (United Nations undated) depends on the availability of comparable data from direct assessments describing the distributions of literacy and numeracy skills among in-school students and adults. In this context, reflecting on issues related to managing data quality in large-scale assessments is important both for understanding the limits of such studies and identifying the ways in which data quality can be improved over time.

Irwin S. Kirsch
Center for Global Assessment, Educational Testing Service, Princeton,
New Jersey, USA
William Thorn
Organisation for Economic Co-operation and Development, Paris, France, and
Matthias von Davier
Center for Advanced Assessment, National Board of Medical Examiners, Philadelphia,
Pennsylvania, USA

References

Groves, R.M. and Lyberg, L. (2010), "Total survey error: past, present, and future", *Public Opinion Quarterly*, Vol. 74 No. 5, pp. 849-879.

Hansen, M., Hurwitz, W. and Madow, W. (1953), *Sample Survey Methods and Theory*, Wiley, New York, NY.

Hansen, M., Hurwitz, W., Marks, E. and Mauldin, P. (1951), "Response errors in surveys", *Journal of the American Statistical Association*, Vol. 46 No. 254, pp. 147-190.

Krosnick, J. (1991), "Response strategies for coping with the cognitive demands of attitude measures in surveys", *Applied Cognitive Psychology*, Vol. 5 No. 3, pp. 213-236, doi: 10.1002/acp.2350050305.

Further reading

United Nations (2018), "Sustainable development goals: 17 goals to transform our world," available at: www.un.org/sustainabledevelopment/education/

About the Guest Editors

Irwin S. Kirsch is the Ralph Tyler Chair in Large Scale Assessment and Director, Center for Global Assessment, at Educational Testing Service in Princeton, New Jersey. In addition to serving as a member of the research management team, his responsibilities include managing and integrating the work of multiple teams consisting of research scientists, data managers, platform developers, research project managers and policy analysts. He also serves as the International Project Director for two key international comparative assessments run through the OECD: PISA (The Programme for International Student Assessment) and PIAAC (The Programme for the International Assessment of Adult Competencies).

William Thorn has managed the OECD's Programme for the International Assessment of Adult Competencies (PIAAC) since August 2007. Prior to joining the OECD, he held a wide range of senior positions in the Australian Federal Departments of Education and Employment. This included the management of units responsible for research into education and the labor market, program evaluation, statistical collections and analysis, tertiary education funding policy and the Commonwealth Government's role in the testing and monitoring of basic skills, such as literacy and numeracy in Australian schools.

Matthias von Davier is a distinguished Research Scientist at the National Board of Medical Examiners. Until 2016, he was a Senior Research Director at Educational Testing Service. He works on modern psychometric methodologies for analyzing data from technology-based high-stakes assessments, is the Co-founding Editor of the journal *Large Scale Assessments in Education*, Editor-in-Chief of the *British Journal of Mathematical and Statistical Psychology* and Co-Editor of a book series *Methodology of Educational Measurement and Assessment*. He received the 2006 ETS Research Scientist award, the 2012 NCME Brad Hanson Award and the 2017 AERA Division D Award for significant contributions to measurement and research methodology.