WILEY SERIES IN SURVEY METHODOLOGY

Implementation of Large-Scale Education Assessments

EDITED BY: Petra Lietz, John C. Cresswell, Keith F. Rust and Raymond J. Adams

WILEY

Implementation of Large-Scale Education Assessments

Wiley Series in Survey Methodology

The Wiley Series in Survey Methodology covers topics of current research and practical interests in survey methodology and sampling. While the emphasis is on application, theoretical discussion is encouraged when it supports a broader understanding of the subject matter.

The authors are leading academics and researchers in methodology and sampling. The readership includes professionals in, and students of, the fields of applied statistics, biostatistics, public policy, and government and corporate enterprises.

ALWIN - Margins of Error: A Study of Reliability in Survey Measurement
BETHLEHEM - Applied Survey Methods: A Statistical Perspective
BIEMER, LEEUW, ECKMAN, EDWARDS, KREUTER, LYBERG, TUCKER, WEST (EDITORS) - Total Survey Error in Practice: Improving Quality in the Era of Big Data
BIEMER - Latent Class Analysis of Survey Error
BIEMER and LYBERG - Introduction to Survey Quality
CALLEGARO, BAKER, BETHLEHEM, GORITZ, KROSNICK, LAVRAKAS (EDITORS) - Online Panel Research: A Data Quality Perspective
CHAMBERS and SKINNER (EDITORS) - Analysis of Survey Data
CONRAD and SCHOBER (EDITORS) - Envisioning the Survey Interview of the Future
COUPER, BAKER, BETHLEHEM, CLARK, MARTIN, NICHOLLS, O'REILLY (EDITORS) - Computer Assisted Survey Information Collection
D'ORAZIO, DI ZIO, SCANU - Statistical Matching: Theory and Practice
FULLER - Sampling Statistics
GROVES, DILLMAN, ELTINGE, LITTLE (EDITORS) - Survey Nonresponse
GROVES, BIEMER, LYBERG, MASSEY, NICHOLLS, WAKSBERG (EDITORS) - Telephone Survey Methodology
GROVES AND COUPER - Nonresponse in Household Interview Surveys
GROVES - Survey Errors and Survey Costs
GROVES - The Collected Works of Robert M. Groves, 6 Book Set
GROVES, FOWLER, COUPER, LEPKOWSKI, SINGER, TOURANGEAU - Survey Methodology, 2nd Edition
HARKNESS, VAN DE VIJVER, MOHLER - Cross-Cultural Survey Methods
HARKNESS, BRAUN, EDWARDS, JOHNSON, LYBERG, MOHLER, PENNELL, SMITH (EDITORS) - Survey Methods in Multicultural, Multinational, and Multiregional Contexts
HEDAYAT, SINHA - Design and Inference in Finite Population Sampling
HUNDEPOOL, DOMINGO-FERRER, FRANCONI, GIESSING, NORDHOLT, SPICER, DE WOLF - Statistical Disclosure Control
KALTON, HEERINGA (EDITORS) - Leslie Kish: Selected Papers
KORN, GRAUBARD - Analysis of Health Surveys
KREUTER (EDITOR) - Improving Surveys with Paradata: Analytic Uses of Process Information
LEPKOWSKI, TUCKER, BRICK, DE LEEUW, JAPEC, LAVRAKAS, LINK, SANGSTER - Advances in Telephone Survey Methodology
LEVY, LEMESHOW - Sampling of Populations: Methods and Applications, 4th Edition
LIETZ, CRESSWELL, RUST, ADAMS (EDITORS) - Implementation of Large-Scale Education Assessments
LUMLEY - Complex Surveys: A Guide to Analysis Using R
LYNN (EDITOR) - Methodology of Longitudinal Surveys
MADANS, MILLER, MAITLAND, WILLIS - Question Evaluation Methods: Contributing to the Science of Data Quality
MAYNARD, HOUTKOOP-STEENSTRA, SCHAEFFER, VAN DER ZOUWEN (EDITORS) - Standardization and Tacit Knowledge: Interaction and Practice in the Survey Interview
MILLER, CHEPP, WILLSON, PADILLA (EDITORS) - Cognitive Interviewing Methodology
PRATESI (EDITOR) - Analysis of Poverty Data by Small Area Estimation
PRESSER, ROTHGEB, COUPER, LESSLER, E. MARTIN, J. MARTIN, SINGER - Methods for Testing and Evaluating Survey Questionnaires
RAO, MOLINA - Small Area Estimation, 2nd Edition
SÄRNDAL, LUNDSTRÖM - Estimation in Surveys with Nonresponse
SARIS, GALLHOFER - Design, Evaluation, and Analysis of Questionnaires for Survey Research, 2nd Edition
SIRKEN, HERRMANN, SCHECHTER, SCHWARZ, TANUR, TOURANGEAU (EDITORS) - Cognition and Survey Research
SNIJKERS, HARALDSEN, JONES, WILLIMACK - Designing and Conducting Business Surveys
STOOP, BILLIET, KOCH, FITZGERALD - Improving Survey Response: Lessons Learned from the European Social Survey
VALLIANT, DORFMAN, ROYALL - Finite Population Sampling and Inference: A Prediction Approach
WALLGREN, A., WALLGREN B. - Register-based Statistics: Statistical Methods for Administrative Data, 2nd Edition
WALLGREN, A., WALLGREN B. - Register-based Statistics: Administrative Data for Statistical Purposes

Implementation of Large-Scale Education Assessments

Edited by

Petra Lietz

John C. Cresswell

Keith F. Rust

Raymond J. Adams

WILEY

This edition first published 2017

Registered Office
John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, United Kingdom

Library of Congress Cataloging-in-Publication Data

Names: Lietz, Petra, editor. | Cresswell, John, 1950– editor. | Rust, Keith, editor. | Adams, Raymond J., 1959– editor.
Title: Implementation of large-scale education assessments / editors, Petra Lietz, John C. Cresswell, Keith F. Rust, Raymond J. Adams.
Other titles: Wiley Series in Survey Methodology
Description: Chichester, UK ; Hoboken, NJ : John Wiley & Sons, 2017. | Series: Wiley Series in Survey Methodology | Includes bibliographical references and index.
Identifiers: LCCN 2016035918 (print) | LCCN 2016050522 (ebook) | ISBN 9781118336090 (cloth) | ISBN 9781118762479 (pdf) | ISBN 9781118762493 (epub)
Subjects: LCSH: Educational tests and measurements.
Classification: LCC LB3051 .L473 2016 (print) | LCC LB3051 (ebook) | DDC 371.26–dc23
LC record available at https://lccn.loc.gov/2016035918

A catalogue record for this book is available from the British Library.

Cover design by Wiley
Cover image: ZaZa Studio/Shutterstock;
(Map) yukipon/Gettyimages

Set in 10/12.5pt Palatino by SPi Global, Pondicherry, India

10 9 8 7 6 5 4 3 2 1

Contents

Notes on Contributors

Raymond J. Adams
Australian Council *for* Educational Research

Alla Berezner
Australian Council *for* Educational Research

Falk Brese
International Association for the Evaluation of Educational Achievement (IEA) Data Processing and Research Center

Mark Cockle
International Association for the Evaluation of Educational Achievement (IEA) Data Processing and Research Center

John C. Cresswell
Australian Council *for* Educational Research

Steve Dept
cApStAn Linguistic Quality Control

Andrea Ferrari
cApStAn Linguistic Quality Control

Eveline Gebhardt
Australian Council *for* Educational Research

Béatrice Halleux
HallStat

Oswald Koussihouèdé
Programme for the Analysis of Education Systems of CONFEMEN (PASEC)

Sheila Krawchuk
Westat

Ema Lagos Campos
Agencia de Calidad de la Educación

Petra Lietz
Australian Council *for* Educational Research

Antoine Marivin
Programme for the Analysis of Education Systems of CONFEMEN (PASEC)

Juliette Mendelovits
Australian Council *for* Educational Research

Christian Monseur
Université de Liège

Dara Ramalingam
Australian Council *for* Educational Research

Keith F. Rust
Westat

Wolfram Schulz
Australian Council *for* Educational Research

Vanessa Sy
Programme for the Analysis of Education Systems of CONFEMEN (PASEC)

Ross Turner
Australian Council *for* Educational Research

Maurice Walker
Australian Council *for* Educational Research

Foreword

The Science of Large-Scale Assessment

Governments throughout the world recognise that the quality of schooling provided to children and young people will be an important determinant of a country's social and economic success in the twenty-first century. In every country, a central question is what governments and school systems can do to ensure that all students are equipped with the knowledge, skills and attributes necessary for effective participation in the future workforce and for productive future citizenship.

To answer this question, countries require quality information, including information on current levels of student achievement, the performances of subgroups of the student population – especially socio-economically disadvantaged students, Indigenous students and new arrivals – and recent trends in achievement levels within a country. Also important is an understanding of how well a nation's schools are performing in comparison with schools elsewhere in the world. Are some school systems producing better outcomes overall? Have some systems achieved superior improvements in achievement levels over time? Are some more effective in ameliorating the influence of socio-economic disadvantage on educational outcomes? Are some doing a better job of developing the kinds of skills and attributes required for life and work in the twenty-first century?

Some 60 years ago, a small group of educational researchers working in a number of countries conceived the idea of collecting data on the impact of countries' educational policies and practices on student outcomes. With naturally occurring differences in countries' school curricula, teaching practices, ways of organising and resourcing schools and methods of preparing and developing teachers and school leaders, they saw the possibility of studying the effectiveness of different educational policies and practices in ways

that would be difficult or impossible in any one country. The cross-national studies that these researchers initiated in the 1960s marked the beginning of large-scale international achievement surveys.

In the decades since the 1960s, international comparative studies of student achievement and the factors underpinning differences in educational performance in different countries have evolved from a research interest of a handful or academics and educational research organisations to a major policy tool of governments across the globe. International surveys now include the OECD's PISA implemented in 75 countries in 2015 and the IEA's Trends in International Mathematics and Science Study implemented in 59 countries in 2015. Other international studies are conducted in areas such as primary school reading, civics and citizenship and ICT literacy. Complementing these international surveys are three significant regional assessment programmes, with a fourth under development. Governments use the results of these large-scale international studies, often alongside results from their own national surveys, to monitor progress in improving quality and equity in school education and to evaluate the effectiveness of system-wide policies and programmes.

The decades since the 1960s have also seen significant advances in methodologies for the planning, implementation and use of international surveys – in effect, the evolution of a science of large-scale assessment.

This book maps an evolving methodology for large-scale educational assessments. Advances in this field have drawn on advances in specific disciplines and areas of practice, including psychometrics, test development, statistics, sampling theory and the use of new technologies of assessment. The book identifies and discusses 13 elements of a complex, integrated science of large-scale assessment – a methodology that begins with a consideration of the policy context and purpose of a study – proceeds through various steps in the design and implementation of a quality assessment programme and culminates in the reporting and dissemination of a study's findings. Each chapter in the book is authored by one or more international authorities with experience in leading the implementation of an element of the described methodology.

As the contributors to this book explain, the science of large-scale assessments is continuing to evolve. The challenges faced by the field and addressed by a number of contributors to this book include the collection of useful, internationally comparable data on a broader range of skills and attributes than have typically been assessed in large-scale surveys. National education systems and governments are increasingly identifying skills and attributes such as collaboration, innovativeness, entrepreneurship and creativity as

important outcomes of school education. The assessment of such attributes may require very different methods of observation and data gathering, including by capitalising on advances in assessment technologies.

An ongoing challenge will be to ensure that the results of large-scale assessments continue to meet their essential purpose: to inform and lead effective educational policies and practices to better prepare all students for life and work in the twenty-first century.

Professor Geoff Masters (AO)
CEO, Australian Council *for* Educational Research (ACER)
Camberwell, Victoria, January 2016

Acknowledgements

The editors gratefully acknowledge the Australian Council *for* Educational Research (ACER), the Australian Department of Foreign Affairs and Trade (DFAT) and Westat for their support of this book.

Particular thanks go to Juliet Young-Thornton for her patient, friendly and effective assistance throughout the process of producing this book.

Abbreviations

ACER	Australian Council *for* Educational Research
ALL	Adult Literacy and Life Skills Survey
ASER	Annual Status of Education Report
BRR	Balanced repeated replication
CBA	Computer-based assessment
CFA	Confirmatory factor analysis
CFI	Comparative fit index
CIVED	Civic Education Study
CONFEMEN	Conference of Education Ministers of Countries using French as the Language of Communication/Conférence des ministres de l'Éducation des Etats et gouvernements de la Francophonie
DIF	Differential item functioning
DPS	Described proficiency scale
EFA	Exploratory factor analysis
ESC	Expected scored curves
ESCS	Economic, social and cultural status
ESS	European Social Survey
ETS	Educational Testing Service
FEGS	Functional Expert Groups
FIMS	First International Mathematics Study
FT	Field trial
ICC	Item characteristic curve
ICCS	International Civic and Citizenship Education Study
ICILS	International Computer and Information Literacy Study
ICT	Information and computer technology
IDB	International database
IDs	Identification variables

IEA	International Association for the Evaluation of Educational Achievement
IIEP	UNESCO International Institute for Educational Planning
ILO	International Labour Organization
IREDU	Institute for Research in the Sociology and Economics of Education
IRM	Item response models
IRT	Item response theory
ISCED	International Standard Classification of Education
ISCO	International Standard Classification of Occupations
ISEI	International Socio-Economic Index of Occupational Status
ITC	International Test Commission
LAMP	Literacy Assessment and Monitoring Programme
LAN	Local area network
LGE	General Education Law/General de Educación
LLECE	Latin American Laboratory for Assessment of the Quality of Education/Laboratorio Latinoamericano de Evaluación de la Calidad de la Educación
LSA	Large-scale assessments
MOS	Measure of size
MS	Main survey
MTEG	Monitoring Trends in Educational Growth
NAEP	United States National Assessment of Educational Progress
NNFI	Non-normed fit index
NPMs	National project managers
OCR	Optical character recognition
OECD	Organisation for Economic Co-operation and Development
PASEC	The Programme for the Analysis of Education Systems of CONFEMEN/Programme d'Analyse des Systèmes Éducatifs de la CONFEMEN
PCA	Principal component analysis
PCM	Partial credit model
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
PL	Parameter logistic model
PPS	Probability proportional to size
PSUs	Primary sampling units
RL	Reading Literacy Study
RMSEA	Root-mean square error of approximation
RP	Response probability

SACMEQ	Southern and Eastern Africa Consortium for Monitoring Educational Quality
SDGs	Sustainable Development Goals
SEA-PLM	Southeast Asian Primary Learning Metrics
SEM	Structural equation modelling
SERCE	Second Regional Comparative and Explanatory Study
SES	Socio-economic status
SIGE	Students General Information System/Sistema Información General de Estudiantes
SIMCE	Sistema de Medición de la Calidad de la Educación
SIMS	Second International Mathematics Study
SISS	Second International Science Study
SITES	Second Information Technology in Education Study
SSUs	Secondary sampling units
TALIS	Teaching and Learning International Survey
TCMAs	Test-Curriculum Matching Analyses
TERCE	Third Regional Comparative and Explanatory Study
TIMSS	Trends in International Mathematics and Science Study
TORCH	Test of Reading Comprehension
TRAPD	Translation, Review, Adjudication, Pretesting, and Documentation
UAENAP	United Arab Emirates (UAE) National Assessment Program
UNESCO	United Nations Educational, Scientific and Cultural Organization
UREALC	UNESCO's Regional Bureau of Education for Latin America and the Caribbean

1

Implementation of Large-Scale Education Assessments

Petra Lietz, John C. Cresswell, Keith F. Rust and Raymond J. Adams

1.1 Introduction

The 60 years that followed a study of mathematics in 12 countries conducted by the International Association for the Evaluation of Educational Achievement (IEA) in 1964 have seen a proliferation of large-scale assessments (LSAs) in education. In a recent systematic review of the impact of LSAs on education policy (Best et al., 2013), it was estimated that LSAs in education are now being undertaken in about 70% of the countries in the world.

The Programme for International Student Assessment (PISA) conducted by the Organisation for Economic Co-operation and Development (OECD) was implemented in 75 countries in 2015 with around 510 000 participating students and their schools. Similarly, the Trends in International Mathematics and Science Study (TIMSS), conducted by the IEA, collected information from schools and students in 59 countries in 2015.

Implementation of Large-Scale Education Assessments, First Edition.
Edited by Petra Lietz, John C. Cresswell, Keith F. Rust and Raymond J. Adams.

This book is about the implementation of LSAs in schools which can be considered to involve 13 key areas. These start with the explication of policy goals and issues, assessment frameworks, test and questionnaire designs, item development, translation and linguistic control as well as sampling. They also cover field operations, technical standards, data collection, coding and management as well as quality assurance measures. Finally, test and questionnaire data have to be scaled and analysed while a database is produced and accompanied by dissemination and the reporting of results. While much of the book has been written from a central coordinating and management perspective, two chapters illustrate the actual implementation of LSAs which highlight the requirements regarding project teams and infrastructure required for participation in such assessments. Figure 1.2 in the concluding section of this chapter provides details regarding where each of these 13 key areas is covered in the chapters of this book.

Participation in these studies, on a continuing basis, is now widespread, as is indicated in Appendix 1.A. Furthermore, their results have become integral to the general public discussion of educational progress and international comparisons in a wide range of countries with the impact of LSAs on education policy being demonstrated (e.g. Baker & LeTendre, 2005; Best et al., 2013; Breakspear, 2012; Gilmore, 2005). Therefore, it seems timely to bring together in one place the collective knowledge of those who routinely conduct these studies, with the aim of informing users of the results as to how such studies are conducted and providing a handbook for future practitioners of current and prospective studies.

While the emphasis throughout the book is on the practical implementation of LSAs, it is grounded in theories of psychometrics, statistics, quality improvement and survey communication. The chapters of this book seek to cover in one place almost every aspect of the design, implementation and analysis of LSAs, (see Figure 1.2), with perhaps greater emphasis on the aspects of implementation than can be found elsewhere. This emphasis is intended to complement other recent texts with related content but which have a greater focus on the analysis of data from LSAs (e.g. Rutkowski, von Davier & Rutkowski, 2013).

This introductory chapter first provides some context in terms of the development of international, regional and national assessments and the policy context in which they occur. Then, the purposes for countries to undertake such assessments, particularly with a view to evidence-based policymaking in education, are discussed. This is followed by a description of the content of the book. The chapter finishes with considerations as to where LSAs might be headed and what is likely to shape their development.

1.2 International, Regional and National Assessment Programmes in Education

The IEA first started a programme of large-scale evaluation studies in education with a pilot study to explore the feasibility of such an endeavour in 1959–1961 (Foshay et al., 1962). After the feasibility study had shown that international comparative studies in education were indeed possible, the first content areas to be tested were mathematics with the First International Mathematics Study conducted by 12 countries in 1962–1967 (Husén, 1967; Postlethwaite, 1967) and the content areas of the six subject surveys, namely, civic education, English as a foreign language, French as a foreign language, literature education, reading, comprehension and science, conducted in 18 countries in 1970–1971. Since then, as can be seen in Appendix 1.A, participation in international studies of education has grown considerably with 59 and 75 countries and economies, respectively, participating in the latest administrations of the TIMSS by the IEA in 2015 and the PISA by the OECD in 2015.

In addition to international studies conducted by the IEA since the late 1950s and by the OECD since 2000, commencing in the mid 1990s, three assessment programmes with a regional focus have been designed and implemented. First, the Conference of Education Ministers of Countries Using French as the Language of Communication (Conférence des ministres de l'Education des États et gouvernements de la Francophonie – CONFEMEN) conducts the Programme d'Analyse des Systèmes Educatifs de la CONFEMEN (PASEC). Since its first data collection in 1991, assessments have been undertaken in over 20 francophone countries not only in Africa but other parts of the world (e.g. Cambodia, Laos and Vietnam). Second, the Southern and Eastern African Consortium for Monitoring Educational Quality (SACMEQ), with the support of the UNESCO International Institute for Educational Planning (IIEP) in Paris, has undertaken four data collections since 1995, with the latest assessment in 2012–2014 (SACMEQ IV) involving 15 countries in Southeast Africa. Third, the Latin-American Laboratory for Assessment of the Quality in Education (LLECE is the Spanish acronym), with the assistance of UNESCO's Regional Bureau for Education in Latin America and the Caribbean (UREALC), has undertaken three rounds of data collection since 1997, with 15 countries participating in the Third Regional Comparative and Explanatory Study (TERCE) in 2013. First steps towards an assessment in the Asia-Pacific region are currently being undertaken through the Southeast Asian Primary Learning Metrics (SEA-PLM) initiative.

In terms of LSAs of student learning, a distinction is made here between LSAs that are intended to be representative of an entire education system,

which may measure and monitor learning outcomes for various subgroups (e.g. by gender or socio-economic background), and large-scale examinations that are usually national in scope and which report or certify individual student's achievement (Kellaghan, Greaney & Murray, 2009). Certifying examinations may be used by education systems to attest achievement at the end of primary or secondary education, for example, or education systems may use examinations to select students and allocate placements for further or specialised study, such as university entrance or scholarship examinations. The focus of this book is on the implementation of LSAs of student learning that are representative of education systems, particularly international assessments that compare education systems and student learning across participating countries.

Parallel to the growth in international assessments, the number of countries around the world administering national assessments in any year has also increased – from 28 in 1995 to 57 in 2006 (Benavot & Tanner, 2007). For economically developing countries in the period from 1959 to 2009, Kamens and Benavot (2011) reported the highest number of national assessments in one year as 37 in 1999. Also in the 1990s, most of the countries in Central and South America introduced national assessments (e.g. Argentina, Bolivia, Brazil, Colombia, Dominican Republic, Ecuador, El Salvador, Guatemala, Paraguay, Peru, Uruguay and Venezuela) through the Partnership for Educational Revitalization in the Americas (PREAL) (Ferrer, 2006) although some introduced them earlier (e.g. Chile in 1982 and Costa Rica in 1986).

International, regional and national assessment programmes can all be considered as LSAs in education. While this book focuses mainly on international assessment programmes conducted in primary and secondary education, it also contains examples and illustrations from regional and national assessments where appropriate.

1.3 Purposes of LSAs in Education

Data from LSAs provide information regarding the extent to which students of a particular age or grade in an education system are learning what is expected in terms of certain content and skills. In addition, they assess differences in achievement levels by subgroups such as gender or region and factors that are correlated with different levels of achievement. Thus, a general purpose of participation in LSAs is to obtain information on a system's educational outcomes and – if questionnaires are administered to obtain background information from students, teachers, parents and/or