

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, KROSINICK, 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
© 2017 by John Wiley and Sons Ltd

Registered Office

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

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com.

The right of the author to be identified as the author of this work has been asserted in accordance with the Copyright, Designs and Patents Act 1988.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, except as permitted by the UK Copyright, Designs and Patents Act 1988, without the prior permission of the publisher.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Designations used by companies to distinguish their products are often claimed as trademarks. All brand names and product names used in this book are trade names, service marks, trademarks or registered trademarks of their respective owners. The publisher is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. It is sold on the understanding that the publisher is not engaged in rendering professional services and neither the publisher nor the author shall be liable for damages arising herefrom. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

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://lcn.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

Contents

Notes on Contributors	xv
Foreword	xvii
Acknowledgements	xx
Abbreviations	xxi
1 Implementation of Large-Scale Education Assessments	1
<i>Petra Lietz, John C. Cresswell, Keith F. Rust and Raymond J. Adams</i>	
1.1 Introduction	1
1.2 International, Regional and National Assessment Programmes in Education	3
1.3 Purposes of LSAs in Education	4
1.3.1 Trend as a Specific Purpose of LSAs in Education	8
1.4 Key Areas for the Implementation of LSAs in Education	10
1.5 Summary and Outlook	16
Appendix 1.A	18
References	22
2 Test Design and Objectives	26
<i>Dara Ramalingam</i>	
2.1 Introduction	26
2.2 PISA	27
2.2.1 Purpose and Guiding Principles	27
2.2.2 Target Population	27
2.2.3 Sampling Approach	28
2.2.4 Assessment Content	29
2.2.5 Test Design	29
2.2.6 Link Items	30

2.3	TIMSS	34
2.3.1	Purpose and Guiding Principles	34
2.3.2	Target Population	34
2.3.3	Sampling Approach	36
2.3.4	Assessment Content	36
2.3.5	Test Design	38
2.4	PIRLS and Pre-PIRLS	41
2.4.1	Assessment Content	41
2.4.2	Test Design	42
2.5	ASER	45
2.5.1	Purpose and Guiding Principles	45
2.5.2	Target Population	46
2.5.3	Sampling Approach	47
2.5.4	Assessment Content	48
2.5.5	Test Design	49
2.6	SACMEQ	52
2.6.1	Purpose and Guiding Principles	52
2.6.2	Target Population	53
2.6.3	Sampling Approach	53
2.6.4	Assessment Content	54
2.6.5	Test Design	55
2.7	Conclusion	56
	References	58
3	Test Development	63
	<i>Juliette Mendelovits</i>	
3.1	Introduction	63
3.2	Developing an Assessment Framework: A Collaborative and Iterative Process	65
3.2.1	What Is an Assessment Framework?	66
3.2.2	Who Should Develop the Framework?	67
3.2.3	Framework Development as an Iterative Process	67
3.3	Generating and Collecting Test Material	68
3.3.1	How Should Assessment Material Be Generated?	69
3.3.2	Who Should Contribute the Material?	69
3.3.3	Processing Contributions of Assessment Material	71
3.4	Refinement of Test Material	72
3.4.1	Panelling of Test Material by Test Developers	73
3.4.2	Panelling Stimulus	73
3.4.3	Panelling Items	74
3.4.4	Cognitive Interviews and Pilot Studies	75

3.4.5	Preparations for Trial Testing	77
3.4.6	Analysis of Trial Test Data	78
3.5	Beyond Professional Test Development: External Qualitative Review of Test Material	81
3.5.1	Jurisdictional Representatives	81
3.5.2	Domain Experts	83
3.5.3	The Commissioning Body	84
3.5.4	Dealing with Diverse Views	84
3.6	Introducing Innovation	86
3.6.1	Case Study 1: The Introduction of Digital Reading in PISA 2009	87
3.6.2	Case Study 2: The Introduction of New Levels of Described Proficiency to PISA in 2009 and 2012	89
3.7	Conclusion	90
	References	90
4	Design, Development and Implementation of Contextual Questionnaires in Large-Scale Assessments	92
	<i>Petra Lietz</i>	
4.1	Introduction	92
4.2	The Role of Questionnaires in LSAs	93
4.3	Steps in Questionnaire Design and Implementation	95
4.3.1	Management of Questionnaire Development Process and Input from Relevant Stakeholders	95
4.3.2	Clarification of Aims and Content Priorities	96
4.3.3	Development of Questionnaires	107
4.3.4	Permissions (Copyright/IP) Requests	109
4.3.5	Cognitive Interviews with Respondents from the Target Population	109
4.3.6	Cultural/Linguistic Adaptations to Questionnaires	111
4.3.7	Ethics Application to Approach Schools and Students	112
4.3.8	Field Trial Questionnaire Administration	112
4.3.9	Analyses of Field Trial Data to Finalise the Questionnaire	113
4.3.10	Selection of Material for the Final MS Questionnaire	114
4.3.11	MS Questionnaire Administration	114
4.3.12	Preparation of Questionnaire Data for Public Release	114
4.4	Questions and Response Options in LSAs	115
4.5	Alternative Item Formats	119

4.6	Computer-Based/Online Questionnaire Instruments	128
4.6.1	Field Trial of Computer-Based Questionnaires	129
4.6.2	Beta Testing	131
4.7	Conclusion and Future Perspectives	131
	Acknowledgements	132
	References	132
5	Sample Design, Weighting, and Calculation of Sampling Variance	137
	<i>Keith F. Rust, Sheila Krawchuk and Christian Monseur</i>	
5.1	Introduction	137
5.2	Target Population	138
5.2.1	Target Population and Data Collection Levels	138
5.2.2	Target Populations of Major Surveys in Education	139
5.2.3	Exclusion	143
5.3	Sample Design	144
5.3.1	Multistage Sample Design	144
5.3.2	Unequal Probabilities of Selection	145
5.3.3	Stratification and School Sample Size	146
5.3.4	School Nonresponse and Replacement Schools	147
5.4	Weighting	148
5.4.1	Reasons for Weighting	148
5.4.2	Components of the Final Student Weight	149
5.4.3	The School Base Weight	150
5.4.4	The School Base Weight Trimming Factor	151
5.4.5	The Within-School Base Weight	151
5.4.6	The School Nonresponse Adjustment	152
5.4.7	The Student Nonresponse Adjustment	152
5.4.8	Trimming the Student Weights	153
5.5	Sampling Adjudication Standards	153
5.5.1	Departures from Standards Arising from Implementation	155
5.6	Estimation of Sampling Variance	156
5.6.1	Introduction	156
5.6.2	Methods of Variance Estimation for Complex Samples	157
5.6.3	Replicated Variance Estimation Procedures for LSA Surveys	158
5.6.4	Computer Software for Variance Estimation	165
5.6.5	Concluding Remarks	165
	References	166

6 Translation and Cultural Appropriateness of Survey Material in Large-Scale Assessments	168
<i>Steve Dept, Andrea Ferrari and Béatrice Halleux</i>	
6.1 Introduction	168
6.2 Overview of Translation/Adaptation and Verification Approaches Used in Current Multilingual Comparative Surveys	169
6.2.1 The Seven Guiding Principles	170
6.2.2 Components from Current Localisation Designs	172
6.3 Step-by-Step Breakdown of a Sophisticated Localisation Design	174
6.3.1 Developing the Source Version(s)	174
6.3.2 Translation/Adaptation	182
6.3.3 Linguistic Quality Control: Verification and Final Check	182
6.4 Measuring the Benefits of a Good Localisation Design	184
6.4.1 A Work in Progress: Proxy Indicators of Translation/Adaptation Quality	186
6.4.2 The Focused MS Localisation Design	187
6.5 Checklist of Requirements for a Robust Localisation Design	190
References	191
7 Quality Assurance	193
<i>John C. Cresswell</i>	
7.1 Introduction	193
7.2 The Development and Agreement of Standardised Implementation Procedures	194
7.3 The Production of Manuals which Reflect Agreed Procedures	196
7.4 The Recruitment and Training of Personnel in Administration and Organisation: Especially the Test Administrator and the School Coordinator	197
7.5 The Quality Monitoring Processes: Recruiting and Training Quality Monitors to Visit National Centres and Schools	198
7.5.1 National Quality Monitors	198
7.5.2 School-Level Quality Monitors	199
7.6 Other Quality Monitoring Procedures	201
7.6.1 Test Administration Session Reports	201
7.6.2 Assessment Review Procedures	202
7.6.3 Checking Print Quality (Optical Check)	202
7.6.4 Post-final Optical Check	202
7.6.5 Data Adjudication Processes	202
7.7 Conclusion	204
Reference	204

8	Processing Responses to Open-Ended Survey Questions	205
	<i>Ross Turner</i>	
8.1	Introduction	205
8.2	The Fundamental Objective	207
8.3	Contextual Factors: Survey Respondents and Items	207
8.4	Administration of the Coding Process	214
8.4.1	Design and Management of a Coding Process	215
8.4.2	Handling Survey Materials	218
8.4.3	Management of Data	218
8.4.4	Recruitment and Training of Coding Personnel	219
8.5	Quality Assurance and Control: Ensuring Consistent and Reliable Coding	221
8.5.1	Achieving and Monitoring Between-Coder Consistency	223
8.5.2	Monitoring Consistency across Different Coding Operations	225
8.6	Conclusion	229
	References	229
9	Computer-Based Delivery of Cognitive Assessment and Questionnaires	231
	<i>Maurice Walker</i>	
9.1	Introduction	231
9.2	Why Implement Computer-Based Assessments?	232
9.2.1	Assessment Framework Coverage	233
9.2.2	Student Motivation	233
9.2.3	Control of Workflow	234
9.2.4	Resource Efficiency	237
9.3	Implementation of International Comparative Computer-Based Assessments	238
9.3.1	Internet Delivery	238
9.3.2	Portable Application	241
9.3.3	Live System	242
9.4	Assessment Architecture	244
9.4.1	Test-Taker Registration	244
9.4.2	Navigation Architecture	245
9.4.3	Assessment Interface	245
9.4.4	Aspect Ratio	247
9.4.5	Accessibility Issues	247
9.5	Item Design Issues	247
9.5.1	Look and Feel	248
9.5.2	Digital Literacy	248
9.5.3	Translation	249

9.6	State-of-the-Art and Emerging Technologies	250
9.7	Summary and Conclusion	250
	References	251
10	Data Management Procedures	253
	<i>Falk Brese and Mark Cockle</i>	
10.1	Introduction	253
10.2	Historical Review: From Data Entry and Data Cleaning to Integration into the Entire Study Process	254
10.3	The Life Cycle of a LSA Study	255
10.4	Standards for Data Management	256
10.5	The Data Management Process	258
10.5.1	Collection of Sampling Frame Information and Sampling Frames	260
10.5.2	School Sample Selection	261
10.5.3	Software or Web-Based Solutions for Student Listing and Tracking	262
10.5.4	Software or Web-Based Solutions for Within-School Listing and Sampling Procedures	263
10.5.5	Adaptation and Documentation of Deviations from International Instruments	265
10.5.6	The Translation Verification Process	266
10.5.7	Data Collection from Respondents	267
10.6	Outlook	272
	References	274
11	Test Implementation in the Field: The Case of PASEC	276
	<i>Oswald Koussihouèdé, Antoine Marivin and Vanessa Sy</i>	
11.1	Introduction	276
11.2	Test Implementation	278
11.2.1	Human Resources	278
11.2.2	Sample Size and Sampling	278
11.2.3	PASEC's Instruments	279
11.2.4	Cultural Adaptation and Linguistic Transposition of the Instruments	289
11.2.5	Preparation of Administrative Documents	289
11.2.6	Document Printing and Supplies Purchase	289
11.2.7	Recruitment of Test Administrators	290
11.2.8	Training, Preparation and Implementation	290

11.2.9	Test Administration	292
11.2.10	Supervision of the Field Work	294
11.2.11	Data Collection Report	294
11.3	Data Entry	294
11.4	Data Cleaning	295
11.5	Data Analysis	295
11.6	Governance and Financial Management of the Assessments	295
	Acknowledgments	296
	References	297
12	Test Implementation in the Field: The Experience of Chile in International Large-Scale Assessments	298
	<i>Emilia Lagos Campos</i>	
12.1	Introduction	298
12.2	International Studies in Chile	302
12.2.1	Human Resources Required in the National Centre	302
12.2.2	Country Input into Instruments and Tests Development	304
12.2.3	Sampling	305
12.2.4	Preparation of Test Materials	307
12.2.5	Preparation and Adaptation of Administrative Documents (Manuals)	309
12.2.6	Preparation of Field Work	310
12.2.7	Actual Field Work	312
12.2.8	Coding Paper and Computer-Based Test	315
12.2.9	Data Entry Process	318
12.2.10	Report Writing	318
12.2.11	Dissemination	320
12.2.12	Final Words	320
	Annex A	321
	References	321
13	Why Large-Scale Assessments Use Scaling and Item Response Theory	323
	<i>Alla Bereznier and Raymond J. Adams</i>	
13.1	Introduction	323
13.2	Item Response Theory	325
13.2.1	Logits and Scales	327
13.2.2	Choosing an IRT Model	328

13.3	Test Development and Construct Validation	329
13.4	Rotated Test Booklets	345
13.5	Comparability of Scales Across Settings and Over Time	347
13.6	Construction of Performance Indicators	349
13.7	Conclusion	354
	References	354
14	Describing Learning Growth	357
	<i>Ross Turner and Raymond J. Adams</i>	
14.1	Background	357
14.2	Terminology: The Elements of a <i>Learning Metric</i>	358
14.3	Example of a Learning Metric	360
14.4	Issues for Consideration	360
14.4.1	Number of Descriptions or Number of Levels	360
14.4.2	Mapping Domain Content onto the Scale	362
14.4.3	Alternative Approaches to Mapping Content to the Metric	363
14.5	PISA Described Proficiency Scales	365
14.5.1	Stage 1: Identifying Scales and Possible Subscales	366
14.5.2	Stage 2: Assigning Items to Subscales	369
14.5.3	Stage 3: Skills Audit	370
14.5.4	Stage 4: Analysing Preliminary Trial Data	371
14.5.5	Stage 5: Describing the Dimension	374
14.5.6	Stage 6: Revising and Refining with Final Survey Data	374
14.6	Defining and Interpreting Proficiency Levels	374
14.7	Use of Learning Metrics	379
	Acknowledgement	380
	References	381
15	Scaling of Questionnaire Data in International Large-Scale Assessments	384
	<i>Wolfram Schulz</i>	
15.1	Introduction	384
15.2	Methodologies for Construct Validation and Scaling	386
15.3	Classical Item Analysis	387
15.4	Exploratory Factor Analysis	388
15.5	Confirmatory Factor Analysis	389
15.6	IRT Scaling	392
15.7	Described IRT Questionnaire Scales	396

15.8	Deriving Composite Measures of Socio-economic Status	399
15.9	Conclusion and Future Perspectives	404
	References	405
16	Database Production for Large-Scale Educational Assessments	411
	<i>Eveline Gebhardt and Alla Berezner</i>	
16.1	Introduction	411
16.2	Data Collection	412
16.3	Cleaning, Recoding and Scaling	416
16.4	Database Construction	418
16.5	Assistance	421
	References	423
17	Dissemination and Reporting	424
	<i>John C. Cresswell</i>	
17.1	Introduction	424
17.2	Frameworks	425
	17.2.1 Assessment Frameworks	425
	17.2.2 Questionnaire Frameworks	426
17.3	Sample Items	426
17.4	Questionnaires	427
17.5	Video	427
17.6	Regional and International Reports	428
17.7	National Reports	428
17.8	Thematic Reports	429
17.9	Summary Reports	429
17.10	Analytical Services and Support	430
17.11	Policy Papers	430
17.12	Web-Based Interactive Display	431
17.13	Capacity-Building Workshops	432
17.14	Manuals	432
17.15	Technical Reports	432
17.16	Conclusion	433
	References	433
	Index	436

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 Koussihouédé

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

Sheila Krawchuk

Westat

Emilia 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 of 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 <i>for</i> 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.

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'Éducation des États et gouvernements de la Francophonie – CONFEMEN) conducts the Programme d'Analyse des Systèmes Éducatifs 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