Chapter 1
Introduction to Reliability and Validity of International Large-Scale Assessment

Hans Wagemaker

Abstract Although international large-scale assessment of education is now a well-established science, non-practitioners and many users often substantially misunderstand how large-scale assessments are conducted, what questions and challenges they are designed to address, and how technologies have evolved to achieve their stated goals. This book focuses on the work of the International Association for the Evaluation of Educational Achievement (IEA), with a particular emphasis on the methodologies and technologies that IEA employs to address issues related to the validity and reliability (quality) of its data. The context in which large-scale assessments operate has changed significantly since the early 1960s when IEA first developed its program of research. The last 60 years has seen an increase in the number of countries participating, with a concomitant expansion in the cultural, socioeconomic, and linguistic heterogeneity of participants. These quantitative and qualitative changes mean that the methodologies and assessment strategies have to evolve continuously to ensure the quality of data is not compromised. This chapter provides an introductory overview of the chronology and development of IEA's international large-scale assessments.

Keywords International Association for the Evaluation of Educational Achievement (IEA) · International large-scale assessment (ILSA) · Large-scale comparative assessment

1.1 Introduction

Since its founding in 1958, IEA has conducted a significant program of research on understanding student achievement in an international context. Its large-scale, comparative studies of education systems are firmly embedded in the educational policy landscape in many countries and sub-national educational jurisdictions. These

H. Wagemaker (✉)
International Association for the Evaluation of Educational Achievement (IEA), Amsterdam, The Netherlands
e-mail: hanswagemaker@compuserve.com

H. Wagemaker (ed.), *Reliability and Validity of International Large-Scale Assessment*, IEA Research for Education 10, https://doi.org/10.1007/978-3-030-53081-5_1

studies of education and schooling have made a significant contribution, not only to advancing understanding of learning outcomes and their antecedents but also to the development of the methodologies that have advanced the science of this field. Furthermore, in keeping with its stated aims, IEA has contributed significantly to the development of the wider research community and to capacity building, particularly with regard to those nations, institutions, and individuals charged with and committed to enhancing the performance of educational systems (Aggarwala 2004; Elley 2005; Gilmore 2005; Lockheed and Wagemaker 2013; Wagemaker 2011, 2013, 2014).

Throughout what is now the considerable history of IEA's international large-scale assessments (ILSAs), a constant concern has been to provide data of the highest possible quality. Researching the complex multi-national context in which IEA studies operate imposes significant burdens and challenges in terms of the methodologies and technologies that are required to achieve the stated study goals to a standard where analysts and policymakers alike can be confident of the interpretations and comparisons that may be made. The demands of the twin imperatives of validity and reliability, tempered by a concern for fairness, must be satisfied in the context of multiple, and diverse cultures, languages, scripts, educational structures, educational histories, and traditions.

Furthermore, as data become more generally accessible and the use of published analyses become more intimately woven into the fabric of the educational reform and policy process, greater attention needs to be paid to more nuanced interpretations of study outcomes.

Each stage of the development and execution of ILSAs, from the framework development that outlines the fundamental research questions to be addressed and the parameters of the study in question (subject, population, and design), to the methodologies used in the conduct of the study (sampling, field operations, quality control, and analysis), to reporting, raise important challenges and questions with respect to ensuring the ultimate reliability and validity of the assessment.

1.2 Outline of This Book

This book offers a comprehensive analysis of the science underpinning all IEA's ILSAs. The content bridges gaps in the general knowledge of consumers of reported outcomes of IEA studies, providing readers with the understanding necessary to properly interpret the results as well as critical insight into the way in which IEA, through its expressed methodologies, has addressed concerns related to quality, with particular reference to issues of reliability and validity.

To a large extent, the chapters that follow reflect the chronology of the development and execution of a large-scale assessment and provide a thorough overview of some of the key challenges that confront developers of IEA assessments and the strategies adopted to address them. While the issues and strategies outlined may also be common to other ILSAs, the focus of this volume is on the work of IEA; the chapters present examples that reflect some of the problems common to

most ILSAs, but include those that are unique to particular IEA studies. As it is not possible to cover every aspect of every study, readers interested in the detail of a particular study are encouraged to consult the related technical reports and documentation that are available on the IEA website (see www.iea.nl).

Following this introductory chapter, Chap. 2 discusses the related concepts of reliability and validity, and provides an overview of the research landscape as it applies to the work of IEA. It reflects on the continued evolution of the research landscape in terms of its magnitude and complexity and how these changes have impacted the process associated with the conduct of IEA's assessments.

The starting point of all assessment is the assessment framework, and Chap. 3 describes the purpose and rationale for developing assessment frameworks for large-scale assessments. As well as differentiating the different types of frameworks, Chap. 3 identifies the steps required for framework development, examining their coverage, publication, and dissemination, and how, procedurally, input from both participants and experts informs their development.

The way in which assessment frameworks are realized in terms of test content is the focus of Chap. 4. Issues related to construct validity, appropriateness of stimulus material in an international context (content), item format, item types, fairness, ability to be translated, sourcing items, and item review processes are considered. Elements like field testing, scorer training, and ultimately, the interpretation of field test outcomes and item adjudication are also addressed.

Chapter 5 embraces the various issues and aspects driving the development of background questionnaires. As for the cognitive tests, item formats, development of indicators, item response coding, measurement of latent constructs, quality control procedures, delivery format options, and future developments are all addressed in this section.

In the multicultural, multilingual, multi-ethnic environment in which ILSAs operate, issues related to translation from the source language to the national test language(s) have fundamental effects on validity and reliability. Chapter 6 addresses the procedures and approaches that IEA has adopted to secure high-quality test instruments in the participants' language of instruction.

For each study, IEA aims to derive reliable and valid population estimates that take into account the complex sample and assessment designs that characterize most studies. Chapter 7 addresses the sampling, weighting, and variance estimation strategies that IEA has developed to take into account and minimize potential sources of error in the calculation of such things as country rankings and subpopulation differences, describing the quality control and adjudication procedures that are key to the construction of validity arguments.

While Chap. 7 focuses on the processes associated with the population definitions and sampling, Chap. 8 focuses on the methodological concerns related to data collection, examining the perspectives of the national study centers for each participating country as well as the views of the international study center. The challenges of monitoring and ensuring data quality are fundamental to ensuring the overall quality of study outcomes and satisfying the imperatives of reliability and validity.

Concerns about data quality are not confined to the methods used to capture data in the field, but extend to the scoring of test booklets and the processing of the questionnaire data. Chapter 9 outlines some of the challenges of quality control, examining issues of comparability and describing the reliability checks that IEA employs to ensure data quality.

New e-technologies for assessment and data collection play an increasingly important role in ILSAs. Chapter 10 examines how those technologies are used to enhance the quality of assessments, their potential for improving measurement and data collection, and the advent of the new knowledge and skill domains related to information technology. In addition to establishing what is best practice in contemporary IEA studies, Chap. 10 explores not only the challenges of developing and delivering technology-based assessments in an international comparative context but also reflects on the potential that e-technologies offer in solving some of the ongoing challenges surrounding ILSA.

Chapter 11 explores the statistical foundations of ILSA and provides an explanation of how the methodologies adopted by the IEA can be used to describe and model individual and system level differences in performance in large-scale surveys of learning outcomes. Beginning with a brief historical overview of the development of latent variable models, this chapter stitches together a critical component of the validity argument as it relates to measurement and analysis in ILSA.

Whereas Chaps. 2 through 11 are intended to address and establish the reliability and validity of the data from a technical point of view, Chaps. 12 through 15 are concerned with aspects related to the issue of consequential validity, namely, the steps taken to enhance impact and mitigate the worst excesses related to misinterpretation, or overinterpretation of IEA's data.

Chapter 12 provides a brief overview of IEA's publication and dissemination strategy, including the quality control procedures for IEA publications. Chapter 13 describes the training and capacity building opportunities that IEA provides to study participants and interested researchers. IEA's investment in skill development aims to provide researchers with sufficient understanding to define good research questions and produce analyses that address issues of interest that are both methodologically and analytically defensible. In keeping with this theme of consequential validity, Chap. 14 provides a case study of best practice from Singapore, demonstrating how data from ILSAs can be used to inform educational policy development and reform.

Finally, Chap. 15 explores the limits and possibilities of using ILSA data to inform that policy process, reminding readers of the need for continual development of the science and proposing a novel way forward in the search for enhancing impact.

References

Aggarwala, N. (2004). *Quality assessment of primary and middle education in mathematics and science (TIMSS).* Evaluation report RAB/01/005/A/01/31. Takoma Park, MD: Eaton-Batson International. https://www.iea.nl/publications/technical-reports/evaluation-report.

Elley, W. B. (2005). How TIMSS-R contributed to education in eighteen developing countries. *Prospects, 35,* 199–212. https://doi.org/10.1007/s11125-005-1822-6.

Gilmore, A. (2005). *The impact of PIRLS (2001) and TIMSS (2003) in low- and middle-income countries: An evaluation of the value of World Bank support for international surveys of reading literacy (PIRLS) and mathematics and science (TIMSS).* Washington, DC: World Bank. https://www.iea.nl/publications/study-reports/international-reports-iea-studies/impact-pirls-2001-and-timss-2003-low.

Lockheed, M. E., & Wagemaker, H. (2013). International large-scale assessments, thermometers, whips or useful policy tools? *Research in Comparative and International Education, 8*(3), 296–306. https://doi.org/10.2304/rcie.2013.8.3.296.

Wagemaker, H. (2011). IEA: International studies, impact and transition. In C. Papanastasiou, T. Plomp, & E. Papanastasiou (Eds.), *IEA 1958–2008: 50 years of experiences and memories* (pp. 253–273). Nicosia, Cyprus: Cultural Center of the Kykkos Monastery. https://www.iea.nl/publications/iea-reference/iea-1958-2008.

Wagemaker, H. (2013). International large-scale assessment (ILSA) programs and the challenges of consequential validity. In M. Chatterji (Ed.), *Validity and test use: An international dialogue on educational assessment, accountability and equity* (pp. 217–233). Bingley, UK: Emerald Publishing.

Wagemaker, H. (2014). International large-scale assessments: From research to policy. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *A handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 11–37). London, UK: Chapman & Hall/CRC Press.