Tertium Comparationis
Journal für International und Interkulturell Vergleichende Erziehungswissenschaft
Vol. 23, No.1, pp. 92–112, 2017

Developments and methodological challenges in international large-scale assessments in education: An IEA perspective

Sabine Meinck, Agnes Stancel-Piątak, Dirk Hastedt & Heiko Sibberns[1]

International Association for the Evaluation of Educational Research (IEA), Hamburg

Abstract

Ongoing societal and technological developments in education and changes in the global debate about education continue to promote the value of international large-scale assessments (ILSAs) in education. ILSAs are expanding their sphere of influence, evolving to cover novel target populations and subject domains. Advances in the methods and technology available to collect, scale, and analyze data present continuous methodological challenges, but also foster rapid developments of the methodological research and respective technology. Most ILSAs in education are now enforcing a transition to computer-based assessment. Recent research has suggested new approaches for addressing nonresponse, novel methods to improve measurement invariance evaluation, and explored innovative methodologies for statistical data analyses. This paper reflects on IEA's extensive experience of ILSA research to identify the most important contemporary challenges, contextualized by historical developments. The authors discuss these developments considering their potentials, drawing conclusions and giving recommendations on best practice.

1. Introduction

The history of international large-scale assessments (ILSAs) of educational provision and achievement now spans more than half a century. The first organization to successfully pursue the idea of empirically exploring the approaches that different educational systems take to educational provision was the International Association for Educational Achievement (IEA). The association's founders wanted to build a body of information that would enable countries to learn from different approaches to education in general and the association between those approaches and student

achievement in particular (IEA, 2016). The first study that IEA conducted was the *Pilot Twelve-Country Study* 1960 (Foshay, Thorndike, Hotyat, Pidgeon & Walker, 1962). Subsequent early studies included the *First International Mathematics Study* (Husén, 1967; Postlethwaite, 1967) and the *First International Science Study* (Bloom, 1969; Comber & Keeves, 1973). Building on that early work, the IEA has developed state-of-art methodology, procedures, and standards that allow reliable crossnational comparisons of the data collected in the many studies the association has conducted since.

The substantial wealth of experience gained from conducting ILSAs now underpins how this and other forms of educational research are conducted, while the data collected offer substantial empirical information on student learning, teaching methods, and school leadership and management for different educational systems all over the world. ILSAs have thus become an internationally accepted tool for monitoring the efficacy of educational systems. However, ongoing societal and technological developments in education and shifts in global debates about education continue to trigger the need for ILSAs to expand their spheres of influence, especially in terms of covering other target populations and additional subject domains. All of these developments, along with advances in the methods and technology available to collect, scale, and analyze data, present ongoing methodological challenges for the teams of researchers conducting ILSAs in education. This article takes a closer look at a variety of former and current developments and the methodological challenges they present. We also consider emerging and likely future developments and challenges. Our exploration is presented from the standpoint of scholars who represent the IEA and have therefore witnessed and contributed to this dynamic process – some of them for more than 20 years. Because these scholars represent the IEA, the perspective provided is that of the IEA.

2. Past and present developments and challenges

2.1 New target populations and domains

When the IEA began developing the field of ILSA research in education in the 1950s and 1960s, it focused primarily on the achievement of students in primary and lower secondary schools, and on just a few subject areas. The association's first study addressed the achievement of primary and lower-secondary school students' achievement in five subject areas: mathematics, reading comprehension, geography, science, and nonverbal ability (Foshay et al., 1962). These subjects were further pursued in later IEA studies, and complemented by other topics such as civic and citizenship education (Oppenheim & Torney, 1974), written composition

(Gorman, Purves & Degenhart, 1988), and the use of computers in education (Pelgrum & Plomp, 1991).

By the mid-1990s, the emphasis in all ILSAs had shifted from covering these diverse areas towards measuring the basic domains of reading, numeracy, and science, steered by the weight placed on accountability and evidence-based education policies. This shift could be seen not only in the emphasis placed on the results of ILSAs but also in the development of national monitoring systems. As the World Bank (n.d.) observed:

> A growing number of governments in the developing world are trying to improve their performance so they can operate more efficiently and provide better services to citizens. To do so, they are creating national or sub-national monitoring and evaluation systems that help them measure and understand how well public programs do. Such systems form the backbone of evidence-based public policy.

In more recent years, ILSA data have also been analyzed in terms of how investment in human capital plays out in countries' productivity and wealth (Rizvi & Lingard, 2010).

Today, IEA's target populations and domains of study are becoming increasingly diverse as more and more education systems recognize that education neither starts before nor finishes after compulsory education, but continues on throughout people's lifetimes and thus encompasses areas of learning and competencies beyond those traditionally taught in schools. This recognition is reflected in Goal 4 of the United Nations (UN) declaration on sustainable development goals: "Ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" (United Nations, 2015, p. 12). Accordingly, the IEA and other agencies, such as the OECD, presently carrying out ILSAs in education are beginning to target pre- and post-compulsory education and are experiencing new challenges with respect to assessment methods, content, and procedures as a consequence. Two examples of these newly investigated areas of educational provision are early childhood education and vocational and technical education (VET).

Early childhood education has become a focus of educational research in many countries as researchers, educators and policymakers increasingly recognize that some of the foundations of inequitable student achievement in primary education are already established as differences in children's abilities and development before they begin school (see, for example, Hart & Risley, 2003). Three recent ILSAs centered on early childhood education are the IEA's *Early Childhood Education Study* (ECES; Bertram & Pascal, 2016), the OECD's *Teaching and Learning International Survey* (TALIS) *Starting Strong Survey* (http://www.iea.nl/oecd-talis-starting-strong-survey), conducted by the IEA and Rand Europe, and the OECD's *International Early Learning and Child Well-being Study* (IELS; http://www.oecd.org/

edu/school/international-early-learning-and-child-well-being-study.htm), conducted by the IEA and the Australian Council for Educational Research (ACER). Conducting large-scale crossnational research into early childhood education is more complex than researching primary and secondary education not only because the structures of the early childhood sector are more diverse within and between countries and the responsibilities for it are less clear within countries but also because the sector is currently undergoing changes and reforms in many countries (Bertram & Pascal, 2016).

Vocational education and training (VET) programs, which students generally enter at the end of their lower secondary schooling, have become an area of greater interest in those countries that have experienced mismatches between educational outcomes and workforce needs (OECD, 2013, p. 24). Countries interested in reforming their VET education are eager to learn from other countries' experiences, and recourse to international comparative data would provide them with a useful starting point (European Commission, 2012; Gill, Fluitman & Dar, 2000). However, the international comparability of this sector is limited due to differences in the structure of VET programs and in the allocation of responsibility for provision. Additional challenges in evaluating VET arise when the private sector also offers education in this area and when the pathways into VET are diverse both across and within countries. VET provision can therefore be dynamic, with changes taking place continuously. This scenario makes it difficult for researchers to establish the baseline data from which they can compare subsequently collected data.

Recognition that the employability of people depends on competencies additional to those of literacy, numeracy, and science has led to policymakers and researchers stressing the important role of education in inculcating social skills (such as the ability to cooperate with other people) for employability (OECD, 2015; Association of Graduate Recruiters, 2017) and for living in a peaceful, sustainable world (UNESCO, 2014). Abilities to work in a team, access new media, use information and communication technologies (ICT), and master foreign languages are a few of the other attributes deemed important in the workplace. However, an increasing number of stakeholders stress that education needs to go further than simply preparing students for the labor market (UNESCO, 2007). There is also the need for education to prepare people for civic participation in today's globalized world, as expressed in Target 4.7 of the UN's fourth sustainable development goal:

> By 2030, ensure that all learners acquire the knowledge and skills needed to promote sustainable development, including, among others, through education for sustainable development and sustainable lifestyles, human rights, gender equality, promotion of a culture of peace and non-violence, global citizenship and appreciation of cultural diversity and of culture's contribution to sustainable development (United Nations, 2015, p. 15).

Over the last two decades, global challenges have posed implications for civic and citizenship education in many countries. Migration has become a pressing political issue, as political and religious persecution and suppression as well as environmental conditions and climate change contribute to burgeoning numbers of refugees. The impact of the global financial crisis of 2007/2008 and the subsequent recession stressed the importance of social cohesion and political stability (Chossudovsky & Marshall, 2010; Grant & Wilson, 2012; Shahin, Woodward & Terzis, 2012) and changed how citizenship is envisioned (Kennedy, 2012), while the rise of ICT and social media is increasingly dominating civic participation. These developments are directing civic and citizenship education perspectives beyond country boundaries and the immediate future towards those of global citizenship education and education for sustainable development (Schulz, Fraillon, Ainley, Losito & Kerr, 2008; UNESCO, 2013, 2015). For the IEA, this global perspective means that the dimensions it measured in its *International Civics and Citizenship Study* (ICCS; http://iccs.iea.nl/) assessments of 2009 and 2016 (with these based on earlier projects conducted in 1971 and 1999) need to be extended. This work will need to identify these competencies and, more importantly, develop assessment items that provide valid and reliable crossnational measures of them. Also, because the IEA's work on civics and citizenship is currently the only established international metric, researchers developing a framework to measure global citizenship competencies will need to rely largely on the IEA's ICCS frameworks.

As countries express greater interest in learning how other countries are providing education in subjects and learning domains additional to those covered in the ILSAs conducted to date, it is likely that these spheres of learning will become part of the ILSA agenda. Singapore, for example, now sends observers to Hungary to learn about their physical education (http://english.tf.hu/tag/forum/). Art and music education, as well as the stronger focus on foreign language learning worldwide, have emerged as topics of crossnational research and discussion, as has geography. Although geography includes natural science content already covered in the IEA's *Trends in Mathematics and Science Study* (TIMSS; http://timssandpirls.bc.edu/), its content is also rooted in the social sciences. These subjects and domains are just some of the others expected to become a focus of future assessment. While those developing ILSAs in these areas will be able to call on the wealth of the ILSA experience developed so far, they will still face methodological challenges peculiar to each discipline.

Another challenge for research teams conducting ILSAs is the increased diversity of participating countries. While mostly high-income countries took part in the earlier assessments, middle- and even low-income countries are now interested in the benefits of ILSA participation (see, for example, the latest TIMSS report;

Mullis, Martin, Foy & Hooper, 2016). The inclusion of these countries increases the diversity of participants in terms of student abilities, a situation that poses challenges for the established measurement capacities of the studies. The IEA first addressed these challenges during the 2011 cycles of TIMSS and PIRLS (the *Progress in Reading and Literacy Study*; http://timssandpirls.bc.edu/) by developing a set of assessment instruments designed specifically to measure student ability at the lower ranges; data collected using these instruments will be translated into the metric of the regular achievement scales in future cycles. Furthermore, the IEA is currently developing a new ILSA, the *Literacy and Numeracy Assessment* (LaNA; www.iea.nl/lana), which is designed to assess basic literacy and numeracy skills and is intended primarily for use in developing education systems. Each system that participates in this study will have opportunity not only to access internationally comparable results (i.e., benchmarking against the established TIMSS and PIRLS achievement scales), thus allowing it to identify the strengths and weaknesses of its educational provision, but also to reap the benefits of capacity building and development of an infrastructure for quantitative research projects. Participation may also provide the foundations of a national education monitoring system, as has happened in many of the countries that have participated in ILSAs.

2.2 The transition from paper-based to computer-based assessment (CBA)

For decades, ILSAs were conducted mostly as paper and pencil tests, where students (or other propositi) received printed survey materials and completed it using a pencil. This situation restricted item formats to multiple-choice questions, short answers, and constructed-response items. It also meant that process information on task completion could not be collected, and that the strategies students used to solve the test or how long it took them to work on particular items remained unknown. National assessment programs, in particular tests conducted in a controlled digital environment (such as, for example, standardized language tests), were first used in the 1990s, but the international assessments continued to be delivered as paper questionnaires. With the advent of personal computers, data-collection procedures began to change and remove the limitations of the paper and pencil method.

Wordprocessing software and desktop publishing programs were among the first computer programs used to expedite ILSAs, starting with the IEA *Reading Literacy Study* conducted in the early 1990s (Lundberg & Linnakylä, 1993). The IEA provided the countries participating in this study with templates of the assessment instruments, a practice which guaranteed that test instruments followed exactly the same layout in each country and represented an important step toward standardized assessment procedures. Clear specification of the type of computer and program in

use – an essential aspect of successful implementation – required a new set of manuals summarizing the technical requirements for participating countries.

The development of data-entry programs swiftly followed, and included the *Data Entry Manager* (DEM), which the IEA developed specifically for its Reading Literacy Study. The DEM not only allowed data to be entered while minimizing various sources of error but also allowed metadata such as variable names, labels, valid ranges, and response category labels to be defined prior to the data-entry process. The use of defined valid ranges meant that entries could easily be verified during the data-capture process, thereby prohibiting the entry of invalid or out-of-range data. Indicators showing column shifts during data entry were established, as were procedures for controlling consistency in the ID system. Export functions and input scripts used the defined metadata information to accommodate a smooth data transfer to standard statistical analysis software. Before 1990, these programs were available only on mainframe computers; subsequently, they could be delivered by personal computer, thus making them instantly usable.

The next development involved programs that supported the within-school sampling process. The first such program developed by IEA was its Within School Sampling Software, used for TIMSS 1995 (Martin & Kelly, 1996). It allowed controlled allocation of different test booklets. It also allowed labels containing ID information to be generated and then attached to the printed test booklet, thus ensuring different respondents were correctly linked (such as teachers with their students) during data collection and also later during data processing. After respondents completed the assessment, the programs were used to track and record the participation status of sampled units and individuals. These programs marked a huge advance in terms of avoiding errors in within-school sampling. They also provided automatic tracking of selection probabilities, which expedited consistency checks between data collected before and during assessment administration.

Fully computerized assessment became possible with the development of tools for digital data collection. At the IEA, this advance proceeded in two steps. First, digital data-collection systems were developed for collecting questionnaire data, the IEA's *Teacher Education and Development Study* (TEDS-M) being the first ILSA to implement this approach. IEA's Online Survey System benchmarked this development, using the internet to transfer respondent data to the study center. The next step was the development of digital assessment systems. The *International Computer and Information Literacy Study* (ICILS; http://www.iea.nl/icils), the IEA's first fully computer-based assessment (CBA), collected both assessment and questionnaire data electronically. This development was quickly followed by an electronic add-on to PIRLS, the ePIRLS assessment of online reading, which completed its data collection in 2016. The OECD's *Programme for International Stu-*

dent Assessment (PISA) moved completely to CBA mode in 2015. The IEA intends to transfer TIMSS to an electronic format in upcoming cycles, while maintaining a paper and pencil option for all education systems where this transition may be challenging in the short term. However, the transition to computerized assessment offers various advantages, such as the opportunity to deliver more complex problem-solving and inquiry tasks, avoid data-entry errors, and reduce the costs that data collection, entry, and scoring incur.

Despite the clear advantages of CBA, computer-assisted data collection has been slow to develop. Aside from the content-related debate concerning the effect of CBA on the measurement of other domains (e.g., McDonald, 2002; Mojarrad, Hemmati, Gohar & Sadeghi, 2013; Sangmeister, 2017), several technical challenges have delayed the development of suitable systems. One constraint concerns translation of the instruments into the languages used in participating countries. Controlled translation and translation verification is vital for ensuring crossnational comparability of ILSA results. The exact meaning of words and phrases needs to be retained in the translation process in order to preserve the difficulty level of the items or the content of questionnaires. Moreover, questionnaire items or answering options sometimes need to be adapted to cultural contexts, or additional items need to be added in order to address specific national research questions. All of these adjustments need to be documented properly in order to make this information available at the data-analysis stage because it is relatively costly to transfer this validation process into an electronic system. During the early days of ILSA, the teams developing the instruments compared paper versions of the original and translated instruments. Later, they were able to track the translation process via a database, but they still had to do a significant amount of manual work. The IEA is now working on a sophisticated and fully computerized system for translation and verification. The advantages of this software will be a seamless process where information, exclusively in electronic formats, is distributed among translators, verifiers, and research coordinators, each of whom will have clearly defined roles. This system will replicate the traditional translation and translation-verification process, but in a digital environment.

The most prominent technical obstacle to computer-based testing in schools is the local availability of computers. If computers are not available, national study centres need to bring laptops (or other devices) to school to conduct the assessment, incurring high costs and demanding logistical procedures. While using school-owned available equipment makes creating a CBA system for a controlled digital environment a fairly straightforward exercise, the digital devices available in schools tend to be extremely heterogeneous, even within countries. They are also subject to security measures defined at the ministry or school level that cannot be

overcome by externally defined software. Various types of operating systems and usage restrictions are in place, and for the newer ILSA studies/study cycles, study administration may need to support not only personal computers or laptops but also tablet devices. To overcome this obstacle, diagnostic programs need to be written to check computers in schools prior to testing. These programs evaluate available disk space, processor speed, working memory, processor type, and screen resolution; they also compare the device's settings against the minimum requirements for the test.

To date, internet-based administration of assessment has yet to be implemented in ILSA, even though survey data collection and upload of assessment materials using the internet is already very common. This possibility will undoubtedly be further explored in the future. At the moment, variations in bandwidth and unease about security measures within participating countries, as well as concerns related to the unauthorized spread of secure test materials, have hindered this development.

We are confident, nonetheless, that CBA is the future of ILSA, with the advantages over paper and pencil assessments being avoidance of cost-intensive printing of assessment materials, immediate availability of data, the development of electronically enhanced item materials, and easier implementation of adaptive testing. Para-data that records the process of test completion at the individual level along with additional criteria (such as response times and behaviors) can provide a deeper understanding of the underlying processes taking place when individuals respond to a question or solve a problem in a test (Goldhammer & Kroehne, 2014; Goldhammer, Naumann & Greiff, 2015). It is likely that recording response patterns (such as speed of response, deletions, repetitions, and changes to the original response), and combining these data with the number of correct responses in order to provide a more informative overall picture of student assessment, will become common practice in LSAs, whether conducted at national or international level.

We caution, however, that the option to use printed materials will still be needed in the future if electronic data collection risks losing participants (whether single respondents or entire groups) and thereby brings in response bias. This issue will probably be most evident when developing countries join a study. Also, costs may constrain the extent to which ILSAs can use digital technologies for data collection and analysis. Careful consideration of possible constraints and challenges is therefore needed in order to judge the feasibility of computer-based ILSA.

2.3 Increasing unit non-participation in large-scale assessments: Impact, consequences, and solutions

The number of studies conducted in education has grown considerably over the last decade. This growth applies to national and international LSAs, as well as to small-

er studies in the educational sector (Kuger, Klieme, Jude & Kaplan, 2016). However, the willingness of schools, students, parents, and teachers to participate in these studies has decreased over the same period. Meinck, Cortes, and Tieck (2017) analyzed the scope of nonresponse in IEA studies over the last decade and found that achieving high participation rates was challenging, especially when adults were targeted; many countries failed to meet the required minimum standards for participation in the studies. Stakeholders identified two major reasons when asked to explain this development. First, schools, school teachers, and parents said they were being 'over-surveyed' (the general perception was that they and their students were being asked to participate in a study too often, with insufficient benefit derived from the time and effort of participation). Second, they expressed criticism of ILSAs and/or other studies in education. Their reasons for this attitude included data-security concerns and ignorance or skepticism about the value, goals, and impact of the studies. Other factors yet to be identified may underlie lack of willingness to participate in studies.

Well-motivated participants result, of course, in high participation rates and reliable response data, both of paramount importance for the quality and validity of the collected data (Armstrong & Overton, 1977; Deming, 1990; Little & Rubin, 1987). Data collection and sampling procedures implemented in ILSA therefore have to fulfill high-quality standards (Gregory & Martin, 2001; OECD, 2014a). Probabilistic random samples are selected before data collection to ensure that study respondents are representative of the populations from which they are drawn, and high participation-rate thresholds are determined to ensure approximate unbiasedness of the results, thereby allowing valid recommendations for policy and practice to be derived from the study results.

In essence, these standards acknowledge the challenges associated with non-participation: the high risk of bias when non-participation is substantial, and – at the same time – non-respondents deviating systematically from respondents with respect to the variables of interest. Because ILSAs have little or no information about the non-respondents at hand, a comprehensive quantification of bias due to non-participation is impossible. As an illustration of the significance of this concern, imagine a situation where, for example, the 20 % of lowest achievers in a given population do not participate in an assessment. In this instance, the average achievement, estimated on the basis of this (biased) sample, would increase considerably. If those reporting these results extrapolated them out to the full population from which the original sample was drawn, the conclusions offered would not be correct and the comparisons made would not be valid.

Participation fatigue and its resultant nonresponse bias have received increasing attention in recent years. The IEA staff and subcontractors[2] involved in conducting

the ICCS and ICILS assessments set up workshops for the studies' national research coordinators (NRCs). During these events, international study coordinators and the NRCs exchanged their ideas and experiences on how to increase participation rates. Some NRCs pointed out that the participation issue is often subject to public debate in their respective countries. Fortunately, the experience IEA has gained from conducting ILSAs over many years has revealed several promising ways to enhance participation rates and evaluate the risk of non-participation bias.

- *Promote the study, taking care to inform all stakeholders and participants adequately:* Participation rates improve if all interested parties are informed in a timely manner about the important aspects of the study.[3] Depending on the audience, aspects such as the aim, value, and significance of the study, the survey operation procedures and timelines, and the measures used to ensure anonymity and data security may be communicated. This information needs to be appropriately tailored for the different audiences (for example, careful selection of illustrations and language). Gaining support from important stakeholders in the field, for example teacher unions, ministries of education, and renowned experts, is also important. Participants should have the opportunity to directly contact their national study center if they have any questions. A good relationship needs to be built between study center staff and the targeted individuals. Recruitment staff should acknowledge they place a burden on participants, and reward their support (even if only orally).
- *Minimize sample overlap between different assessments:* Good coordination across the teams involved in school sampling for each study can help to avoid selecting schools for more than one study. Usually, samples are selected not by experts from within countries but by experts appointed at the international level. Today, the three big players in ILSA sampling – Westat, Statistics Canada, and the IEA – are increasingly cooperating closely in order to control for sample overlap across the different large-scale assessments. The respective approaches are introduced to participating countries, and are often offered by default. If participating countries request overlap control, the sampling teams will accommodate the request as long as their doing so retains sound sampling methodology. Overlap control is also being employed at the national level. Germany and the United States, for example, have controls in place to avoid sample overlap between national assessments and ILSAs.
- *Keep the burden for participants as low as possible:* Efforts to prepare the data collection within, for example, schools should be dealt with as much as possible by the study center, instead of burdening school staff with this work. Support can be offered for filling listing forms and for scheduling and conducting the as-

sessment. During instrument development, assessment and questionnaire experts should carefully balance research interests against feasibility concerns. While researchers may wish to include a broad variety of questions into the instruments so as to cover a wide range of domains, questionnaires must not exceed reasonable lengths. If they are too long, respondents may be unwilling or unable to complete them. The same concern applies to assessments. Rotated or adaptive designs make it possible to cover broad domains while keeping the testing/survey time at a minimum. Engaging materials can be used to retain participants' attention, while digital environments may allow enhanced item and questionnaire formats. Translated materials can be made available to respondents who are not familiar with the official language (for example, those with parents from immigrant backgrounds), as this will help avoid minority-group response bias.

- *Reward participation:* Participation in a study naturally involves effort from the participants. Rewarding this effort can help to increase participation rates. What does or does not work as an incentive varies substantially across countries. Habits, culture, and laws have to be considered when determining the rewards for participation. Examples of successful incentives include, amongst others, handing out small presents, donating money or sponsoring events at schools, paying (adult) participants, letting respondents participate in a lottery, and providing educational development courses.
- *Give feedback on study results:* Relevant feedback on study results can also be an important participation incentive for schools. For this reason, many countries provide various ways of providing feedback particular to a school or a group of schools (Gandal & McGiffert, 2003; Gray, 2002; Rolff, 2002; von der Gathen, 2011). However, the national study centers conducting each ILSA rarely actively support this incentive method. The reasons why include the following. First, ILSA study designs are generally unsuitable for the derivation of individual or small group estimators, but are instead optimized for evaluations at the system level (Mirazchiyski, 2013). Second, substantive changes in the design of the study would be needed to overcome this obstacle. Third, tightly clocked process sequences and strict regulations with respect to the confidentiality of the data, especially prior to publication, impede the incorporation of feedback systems (Meinck, 2016). However, feedback to schools can be a suitable method of enhancing participation if the feedback respects the limits imposed by study designs. Feedback should be framed in a way that brings out the implications of the core study findings for the daily work of practitioners and principals, and it can take several forms, ranging from paper reports and professional development courses to conferences tailored to the appropriate audience.

- *Consider (new) approaches to analyzing non-participation bias risks:* At present, standard procedures for evaluating this risk in ILSAs have yet to be developed. Instead, nonresponse adjustments are conducted under the assumption of non-informative response models within sampling strata (for an overview, see, for example, Martin & Mullis, 2012; Meinck, 2015; Meinck & Cortes, 2015). According to Meinck et al. (2017), only three large international comparative surveys in education currently conduct systematic nonresponse bias analysis in an effort to evaluate the risk of bias due to poor participation. They are the OECD's *Programme for the International Assessment of Adult Competencies* (PIAAC; http://www.oecd.org/skills/piaac/) (Mohadjer, Krenzke & Van de Kerckhove, 2013), the OECD's *Teaching and Learning International Survey* (TALIS; https://www.oecd.org/edu/school/talis.htm) (OECD, 2014c), and the IEA's *International Civic and Citizenship Study* (ICCS; Meinck & Cortes, 2015). However, Meinck and her colleagues consider the scope and validity of these analyses relatively limited, and therefore propose (Meinck et al., 2017) use of a 'school nonresponse questionnaire' (a shortened school questionnaire), designed to inform the nonresponse adjustments. This approach has yet to be trialed. Further research on similar methodological approaches is also needed to expand the methodological toolbox for addressing the issues associated with nonresponse.

3. New possibilities and challenges for large-scale data analysis

During the last two decades, ongoing developments in ICT have been continually enhancing the possibilities for data mining in the fields of econometrics, sociology, psychology, market research, astronomy, oceanography, and engineering (Halevi & Moed, 2012). Within the field of educational ILSA, the evolution of ICT has revolutionized data-collection (with e-assessments gradually replacing paper and pencil tests), data-analysis and data-validation techniques and so enhanced the possibilities for including a growing number of participants (whether individuals or groups/countries) in these large-scale international studies. Today, the global education research community has available to them not only 60 years of ILSA data, but also tools for analyzing this empirical information that have been enhanced, particularly within the last 20 years, by the developments in ICT and through effort to align the surveys across studies and across study cycles (for example, PIRLS and TIMSS).

These large data sets are providing more information (more items) about a larger number of subjects than ever before, while ICT developments are enabling greater accessibility to that information. User-friendly data-analysis software now allows

less technically affiliated researchers to run the more complex and comprehensive analyses themselves. In terms of methodology, these developments mean that ILSA researchers can now collect more comprehensive and novel kinds of data (for example, electronic information on response time and patterns in e-based assessments) and obtain more precise and comprehensive information from that data.

As the number of measurement points and objects (cycles and countries) grows and the accessibility of user-friendly analytical tools improves (see, for example, Mplus, https://www.statmodel.com/; HLM, http://www.ssicentral.com/hlm/index.html; LISREL, http://www.ssicentral.com/lisrel/index.html), methods such as multilevel regression modeling or structural equation modeling are becoming increasingly widespread in educational research. These complex methods have the advantage of being more suitable for mirroring the increasing complexity of theoretical frameworks. In addition, the growing amount of data and the greater availability of technological tools (computer-processing capacity and computer software) are allowing researchers and other stakeholders to obtain more specific information about subgroups, such as different minority groups within countries, than has previously been possible. In the past, the more limited scale of data collections made it difficult to obtain sufficient information about these groups.

Enhanced data-collection and data-processing techniques are also making it easier to collect data more efficiently and reliably for selected subpopulations within countries. Although past research teams were well aware of crosscultural differences across subpopulations, they found identifying differences with respect to, for instance, test validity across cultures challenging because of limited computer capacities, non-availability of user-friendly software, and insufficient knowledge of the applications and of suitable data-analysis techniques. This situation is now being turned around through enhanced computer capacity, the availability of user-friendly software, and more ready access to knowledge (e.g., online tutorials, internet forums). These developments are also making it easier to analyze, for example, the validity of test items or of psychological and sociological constructs for the greater number of populations and subgroups wanting to participate in ILSAs. All of these recently developed techniques and procedures are now being evaluated and gradually incorporated into standard procedures in ILSA.

At present, it is still easier to achieve crosscultural comparability of psychological or sociological constructs (i.e., ensuring measurement invariance; Meredith, 1993; Rutkowski & Svetina, 2013) in those ILSAs involving smaller numbers of countries (see, for example, Stancel-Piątak & Desa, 2014). Research conducted over the last 15 or so years illustrates the types of difficulties that ILSAs involving large numbers of countries have experienced when trying to ensure measurement invariance (OECD, 2010, 2014b; Schulz, 2009; Schulz & Friedman, 2011; Schulz

& Sibberns, 2004). The problem of measurement invariance has taxed methodologists for much longer, however, with various alternative procedures having been developed over the last few decades in an effort to validate crosscultural comparability of constructs. These methods include *partial invariance* (Byrne, Shavelson, & Muthén, 1989) and *approximate invariance* (Asparouhov & Muthén, 2014; Marsh, Liem, Martin, Morin & Nagengast, 2011; Muthén & Asparouhov, 2013; Van de Schoot, Klutymans, Tummers, Lugtig, Hox & Muthén, 2013; Van de Schoot, Schmidt, De Beuckelaer, Lek & Zondervan-Zwijnenburg, 2015). The assumption underlying these methods is that the latent construct will be very similar across countries, but not identical. Future research will reveal if crosscultural comparisons can benefit from this more realistic assumption and if the respective analysis techniques will allow us to draw greater knowledge from ILSA data.

Predictive models developed under the umbrella of machine science (also called machine learning; Essa & Ayad, 2012; Kaplan & Lee, 2015) are also beginning to feature among the tools being used to bring greater precision and validity to the analysis of ILSA data. Initially developed within the research field of artificial intelligence with the aim of enhancing computers' capacity to 'learn' from experience, machine learning has since been adapted for use in different disciplines, ranging from economics through healthcare, weather forecasting, and market research, to archeology. Predictive models are one such adaptation. These models employ machine intelligence statistical probability techniques to mine empirical data for variables likely to aid prediction of future outcomes. These outcomes generally concern unknown events in the future, but are not necessarily limited to future events. Predictive modeling in education can also be used, for instance, to identify at-risk students pre-emptively (Essa & Ayad, 2012) or to drive effective interventions (Smith, Lange & Huston, 2012). Within the context of ILSAs, student outcomes at the country level can be predicted in a more reliable way through the use of data from multiple cycles (Huang, 2011; Kaplan & Lee, 2015). They can additionally be used to gain insights into hidden patterns, identified from previous learning relating to historical relationships and trends.

While critically discussed and criticized for being a shortcut not well-founded in a theoretical framework, predictive models have some advantages compared to traditional methods. Advocates of predictive models argue that they can contribute to hypothesis generation, development of new measures, and indicate the potential level of model predictability (Shmueli, 2010, p. 292). Advocates also claim that these models can bring to the fore findings that ultimately are more reliable and repeatable than are the findings derived from classical analysis methods. Predictive modeling also makes it possible to combine different models, such as decision trees and neural network analysis, with regression analysis. Another advantage is the

opportunity to use different measures of model fit – a need that arises from the fact that the *p*-value is sample-size sensitive, resulting in a significant *p*-value for all parameter estimates calculated using large data sets. This method also robustly handles multicollinearity, as the aim is not to estimate individual parameters, but rather to use all available information for the best prediction (ibid.).

Because ILSAs in education still focus mainly on school surveys and may be conducted in developing countries where the technological resources for e-assessments are limited, methodological developments and innovations are not occurring as quickly in these studies as they are in other areas such as econometrics or market research. ILSAs in education employing cyclical data collection and standardized or aligned tests and questionnaires (such as TIMSS and PIRLS) demand the production of large data sets for comparative analysis, and will increasingly continue to do so in the future. At the same time, the increasing amount of information will continue to create challenges, one such being protection of personal data. Another challenge is the difficulty individuals who have a stake in the studies but do not have technical expertise are likely to experience when trying to access and draw information from the available data sets. Although the rapid development of modern visualization methods means that ILSA results can be presented in a straightforward manner and easily understood by policymakers, practitioners, and other stakeholders, the accessibility concern is one that the ILSA research community needs to address sooner rather than later. Finally, new analysis methods such as the predictive modeling need to be subject to ongoing discussion and critical reflection on their suitability for the theoretical framework underpinning each ILSA, as well as the value they add to research, education, and practice.

4. Final remarks

Even though international assessments are now well-established monitoring instruments in education, they are continuing to develop and evolve. The recently defined or refined agendas of international stakeholders in education (such as the United Nations' sustainable development goals for 2030) mean that ILSAs are now drawing in new populations and covering additional subjects. Today, many stakeholders accept that education neither starts at primary school nor ends after secondary school, but is a lifelong endeavor. Therefore, future ILSAs will likely focus on early childhood education, vocational education, and lifelong learning. Stakeholders also agree that mathematics and reading skills are base competencies, but need to be complemented by other skills, such as information and computer literacy, or

social and language skills, topics that have either emerged recently on the agenda of comparative assessment or are the subject of future proposed ILSAs.

Developments such as these will continue to throw up methodological challenges for the teams designing and implementing these large-scale crossnational studies and analyzing the data collected. However, equally rapid developments in methodology, including those expedited by information and communication technologies, are helping research teams address these challenges. Most ILSAs in education are now enforcing a transition to computer-based assessment; new approaches for addressing nonresponse are being discussed; processes to improve measurement invariance evaluation are being designed; and innovative methodologies of statistical data analysis are being explored. The collaborative spirit within the larger ILSA research community makes us confident that innovative methodological techniques and tools will help us keep pace with the challenges associated with this rapidly developing field of educational research in the coming years.

Notes

1. We thank Gillian Wilson (IEA Secretariat, Amsterdam) and Paula Wagemaker for their thorough editorial review.
2. Staff at the Australian Council for Education Research (ACER).
3. Source: Personal talks with NRCs about the reasons for increased participation rates in specific studies across cycles.

References

Armstrong, J.S. & Overton, T.S. (1977). Estimating nonresponse bias in mail surveys. *Journal of Marketing Research*, *14* (3), 396–402.

Asparouhov, T. & Muthén, B. (2014). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, *21* (4), 495–508. doi:10.1080/10705511.2014.919210

Association of Graduate Recruiters. (2017). *AGR Development Survey 2017*. Retrieved March 22, 2017, from http://www.agr.org.uk/AGR-Development-Survey-2017.

Bertram, T. & Pascal, C. (2016). *Early childhood policies and systems in eight countries: Findings from IEA's early childhood education study*. Amsterdam: International Association for the Evaluation of Educational Achievement.

Bloom, B.S. (1969). *Cross-national study of educational attainment: Stage I of the I.E.A. investigation in six subject areas* (Vols. 1–2). Washington, DC: Office of Education (DHEW).

Byrne, B.M., Shavelson, R.J. & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, *105* (3), 456–466. doi:10.1037/0033-2909.105.3.456

Chossudovsky, M. & Marshall, A.G. (Eds.). (2010). *The global economic crisis: The great depression of the XXI century*. Montréal: Global Research Publishers.

Comber, L.C. & Keeves, J.P. (1973). *Science education in nineteen countries: An empirical study*. Stockholm: Almqvist & Wiksell.

Deming, W.E. (1990). *Sample design in business research*. New York: Wiley & Sons.

Essa, A. & Ayad, H. (2012). Improving student success using predictive models and data visualisations. *Research in Learning Technology*, *20* (supplemental issue). Retrieved March 22, 2017, from http://www.tandfonline.com/doi/full/10.3402/rlt.v20i0.19191

European Commission. (2012). *TVET and skills development in EU development cooperation. 2012/308055/1. Final report*. Retrieved March 22, 2017, from http://ec.europa.eu/europeaid/skills-and-vocational-education-trainings-eu-development-cooperation-document_en

Foshay, A.W., Thorndike, R.L., Hotyat, F., Pidgeon, D.A. & Walker, D.A. (1962). *Educational achievements of thirteen-year-olds in twelve countries: Results of an international research project, 1959–1961*. Hamburg: UNESCO Institute for Education.

Gandal, M. & McGiffert, L. (2003). The power of testing. *Educational Leadership*, *60* (5), 39–42.

Gill, I.S., Fluitman, F. & Dar, A. (Eds.). (2000). *Vocational education and training reform: Matching skills to markets and budgets*. Washington, DC: Oxford University Press.

Goldhammer, F. & Kroehne, U. (2014). Controlling individuals' time spent on task in speeded performance measures: Experimental time limits, posterior time limits, and response time modeling. *Applied Psychological Measurement*, *38* (4), 255–267.

Goldhammer, F., Naumann, J. & Greiff, S. (2015). More is not always better: The relation between item response and item response time in Raven's matrices. *Journal of Intelligence*, *3*, 21–40.

Gorman, T.P., Purves, A.C. & Degenhart, R.E. (Eds.). (1988). *The IEA study of written composition I: The international writing tasks and scoring scales*. Oxford: Pergamon Press.

Grant, W. & Wilson, G.K. (Eds.). (2012). *The consequences of the global financial crisis: The rhetoric of reform and regulation*. Oxford: Oxford University Press.

Gray, J. (2002). Jolts and reactions: Two decades of feeding back information on schools' performance. In A.J. Visscher & R. Coe (Eds.), *School improvement through performance feedback* (pp. 142–162). Lisse: Swets & Zeitlinger.

Gregory, K.D. & Martin, M.O. (2001). *Technical standards for IEA studies: An annotated bibliography*. Amsterdam: International Association for the Evaluation of Educational Achievement.

Halevi, G. & Moed, H.F. (2012). The evolution of big data as a research and scientific topic: Overview of the literature. *Research Trends*, *30*. Retrieved March 22, 2017, from https://www.researchtrends.com/issue-30-september-2012/the-evolution-of-big-data-as-a-research- and-scientific-topic-overview-of-the-literature/

Hart, B. & Risley, T.R. (2003). The early catastrophe: The 30 million word gap by age 3. *American Educator*, *27* (1), 4–9.

Huang, S. (2011). *Predictive modeling and analysis of student academic performance in an engineering dynamics course*. Logan, UT: Utah State University. Retrieved March 22, 2017, from http://digitalcommons.usu.edu/cgi/viewcontent.cgi?article=2097&context=etd

Husén, T. (Ed.). (1967). *International study of achievement in mathematics: A comparison of twelve countries* (Vols. 1–2). Stockholm: Almqvist & Wiksell.

IEA (International Association for the Evaluation of Educational Achievement). (2016). *Brief history of the IEA*. Retrieved December 7, 2016, from http://www.iea.nl/brief-history-iea

Kaplan, D. & Lee, C. (2015). Bayesian model averaging over directed acyclic graphs with implications for the predictive performance of structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, *23* (3), 343–353. doi:10.1080/10705511.2015.1092088

Kennedy, K.J. (2012). Global trends in civic and citizenship education: What are the lessons for nation states? *Education Sciences*, *2* (3), 121–135. doi:10.3390/educsci2030121

Kuger, S., Klieme, E., Jude, N. & Kaplan, D. (Eds.). (2016). *Assessing contexts of learning: An international perspective*. New York: Springer.

Little, R.J.A. & Rubin, D.B. (1987). *Statistical analysis with missing data*. New York: Wiley.

Lundberg, I. & Linnakylä, P. (1993). *Teaching reading around the world: IEA Study of Reading Literacy*. The Hague: IEA.

Marsh, H., Liem, G., Martin, A., Morin, A. & Nagengast, B. (2011). Methodological measurement fruitfulness of exploratory structural equation modeling (ESEM): New approaches to key substantive issues in motivation and engagement. *Journal of Psychoeducational Assessment*, *29* (4), 322–346. doi:10.1177/0734282911406657

Martin, M.O. & Kelly, D.L. (Eds.). (1996). *TIMSS technical report: Vol. I. Design and development*. Chestnut Hill, MA: Boston College.

Martin, M.O. & Mullis, I.V.S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: Boston College.

McDonald, A.S. (2002). The equivalence of computer-based and paper-and-pencil educational assessments. *Computers & Education, 39*, 299–312.

Meinck, S. (2015). Computing sampling weights in large-scale assessments in education. *Survey Methodes: Insights from the Field*. Retrieved March 22, 2017, from http://surveyinsights.org/?p=5353

Meinck, S. (2016). Rückmeldung an Schulen im Rahmen von Large-Scale Assessments: Chancen und Grenzen aus methodischer Sicht. Paper presented at the Kommission Bildungsplanung, Bildungsorganisation und Bildungsrecht (KBBB), 29. September, Paderborn.

Meinck, S. & Cortes, D. (2015). Sampling weights, nonresponse adjustments and participation rates. In J. Fraillon, W. Schulz, T. Friedman, J. Ainley & E. Gebhardt (Eds.), *International Computer and Information Literacy Study 2013 technical report* (pp. 87–111). Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

Meinck, S., Cortes, D. & Tieck, S. (2017). Evaluating the risk of nonresponse bias in educational large-scale assessments with school nonresponse questionnaires: A theoretical study. *Large-scale Assessments in Education*, *5* (3). Retrieved March 22, 2017, from http://rdcu.be/oVhH

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58* (4), 525–543.

Mirazchiyski, P. (2013). *Providing school-level reports from international large-scale assessments: Methodological considerations, limitations, and possible solutions*. Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

Mohadjer, L., Krenzke, T. & Van de Kerckhove, W. (2013). Indicators of the quality of the sample data. In I. Kirsch & W. Thorn (Eds.), *Technical report of the Survey of Adult Skills (PIAAC)* (chapter 16, pp. 1–30). Paris: OECD Publishing.

Mojarrad, H., Hemmati, F., Gohar, M.J. & Sadeghi, A. (2013). Computer-based assessment (CBA) vs. paper/pencil-based assessment (PPBA): An investigation into the performance and attitude of Iranian EFL learners' reading comprehension. *International Journal of Language Learning and Applied Linguistics World, 4* (4), 418–428.

Mullis, I.V.S., Martin, M.O., Foy, P. & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Retrieved March 22, 2017, from http://timssandpirls.bc.edu/timss2015/international-results/

Muthén, B. & Asparouhov, T. (2013). *New methods for the study of measurement invariance with many groups*. Retrieved March 22, 2017, from http://www.statmodel.com/download/PolAn.pdf

OECD (Organisation for Economic Co-operation and Development). (2010). *TALIS 2008 technical report*. Paris: OECD Publishing. Retrieved March 22, 2017, from http://www.oecd.org/education/school/44978960.pdf

OECD (Organisation for Economic Co-operation and Development). (2013). *Private sector development policy handbook: Developing skills in Central Asia through better vocational education and training systems*. Retrieved March 22, 2017, from https://www.oecd.org/globalrelations/VocationalEducation.pdf

OECD (Organisation for Economic Co-operation and Development). (2014a). *PISA 2012 technical report*. Paris: OECD Publishing. Retrieved March 22, 2017, from http://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf

OECD (Organisation for Economic Co-operation and Development). (2014b). *TALIS 2013 results: An international perspective on teaching and learning*. Paris: OECD Publishing. doi:10.1787/9789264196261-en

OECD (Organisation for Economic Co-operation and Development). (2014c). *TALIS 2013 technical report*. Paris: OECD Publishing. Retrieved March 22, 2017, from http://www.oecd.org/edu/school/TALIS-technical-report-2013.pdf

OECD (Organisation for Economic Co-operation and Development). (2015). *OECD skills outlook 2015: Youth, skills and employability*. Paris: OECD Publishing. Retrieved March 22, 2017, from http://www.oecd-ilibrary.org/content/book/9789264234178-en

Oppenheim, A.N. & Torney, J. (1974). *The measurement of children's civic attitudes in different nations*. Stockholm: Almqvist & Wiksell.

Pelgrum, W.J. & Plomp, T. (1991). *The use of computers in education worldwide: Results from the IEA 'Computers in Education' survey in 19 educational systems*. Oxford: Pergamon Press.

Postlethwaite, T.N. (1967). *School organization and student achievement: A study based on achievement in mathematics in twelve countries*. Stockholm: Almqvist & Wiksell.

Rizvi, F. & Lingard, B. (2010). *Globalizing education policy*. New York: Routledge.

Rolff, H.-G. (2002). Rückmeldung und Nutzung der Ergebnisse von großflächigen Leistungsuntersuchungen. Grenzen und Chancen. In H.-G. Rolff, H.G. Holtappels, K. Klemm, H. Pfeiffer, & R. Schulz-Zander (Eds.), *Jahrbuch der Schulentwicklung, Bd. 12. Daten, Beispiele und Perspektiven* (pp. 75–98). Weinheim: Juventa.

Rutkowski, L. & Svetina, D. (2013). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*, 31–57.

Sangmeister, J. (2017). Commercial competence: Comparing test results of paper-and-pencil versus computer-based assessments. *Empirical Research in Vocational Education and Training, 9* (3). doi:10.1186/s40461-017-0047-2

Schulz, W. (2009). Questionnaire construct validation in the International Civic and Citizenship Education Study. *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, *2*, 113–136.

Schulz, W., Fraillon, J., Ainley, J., Losito, B. & Kerr, D. (2008). *International Civic and Citizenship Education Study: Assessment framework*. Amsterdam: IEA.

Schulz, W. & Friedman, T. (2011). Scaling procedures for ICCS questionnaire items. In W. Schulz, J. Ainley, & J. Fraillon (Eds.), *ICCS 2009 technical report* (pp. 157–259). Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

Schulz, W. & Sibberns, H. (2004). Scaling procedures for cognitive items. In W. Schulz & H. Sibberns (Eds.), *IEA Civic Education Study technical report* (pp. 69–91). Amsterdam: International Association for the Evaluation of Educational Achievement (IEA).

Shahin, J., Woodward, A. & Terzis, G. (2012). *Study on the impact of the crisis on civil society organizations in the EU: Risks and opportunities*. Belgium: Institute for European Studies, Vrije Universiteit Brussels.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25* (3), 289–310.

Smith, V.C., Lange, A. & Huston, D.R. (2012). Predictive modeling to forecast student outcomes and drive effective interventions in online community college courses. *Journal of Asynchronous Learning Networks*, *16* (3), 51–61.

Stancel-Piątak, A. & Desa, D. (2014). Methodological implementation of multi group multilevel SEM with PIRLS 2011: Improving reading achievement. In R. Strietholt, W. Bos, J.-E. Gustafsson & M. Rosén (Eds.), *Educational policy evaluation through international comparative assessments* (pp. 75–93). Münster: Waxmann.

UNESCO. (2007). *A human rights-based approach to education for all*. New York: United Nations Educational, Scientific and Cultural Organization.

UNESCO. (2013). *Global citizenship education: An emerging perspective*. Retrieved March 22, 2017, from http://www.unesco.org/new/en/gefi/resources/gced/

UNESCO. (2014). *Global citizenship education: Preparing learners for the challenges of the twenty-first century*. Paris: United Nations Educational, Scientific and Cultural Organization.

UNESCO. (2015). *Global citizenship education: Topics and learning objectives*. Paris: United Nations Educational, Scientific and Cultural Organization.

United Nations. (2015). *Transforming our world: The 2030 agenda for sustainable development*. New York: United Nations General Assembly.

Van de Schoot, R., Kluytmans, A., Tummers, L., Lugtig, P., Hox, J. & Muthén, B. (2013). Facing off with Scylla and Charybdis: A comparison of scalar, partial, and the novel possibility of approximate measurement invariance. *Quantitative Psychology and Measurement*, *4*, 770. doi:10.3389/fpsyg.2013.00770

Van de Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K. & Zondervan-Zwijnenburg, M. (2015). Editorial: Measurement invariance. *Frontiers in Psychology*, *6*. doi:10.3389/fpsyg.2015.01064

von der Gathen, J. (2011). *Leistungsrückmeldungen bei Large-Scale Assessments und Vollerhebungen. Rezeption und Nutzung am Beispiel von DESI und lernstand* (Internationale Hochschulschriften, Bd. 552). Münster: Waxmann.

World Bank. (n.d.). Evidence based public policy: Overview. New York: World Bank. Retrieved December 14, 2016, from http://www.worldbank.org/en/topic/evidencebasedpublicpolicy/overview#1