National Academy of Education

Workshop Series on Methods and Policy Uses of International Large-Scale Assessments (ILSAs)

Questionnaire Development and Design for International Large-Scale Assessments (ILSAs): Current Practice, Challenges, and Recommendations

Nina Jude and Susanne Kuger
Educational Quality and Evaluation
The German Institute for International Educational Research

2018

Contact:
Nina Jude, Educational Quality and Evaluation, The German Institute for International Educational Research, Schloßstraße 29, 60486 Frankfurt am Main, +49 (0)69-24708-111, jude@dipf.de

Susanne Kuger, Educational Quality and Evaluation, The German Institute for International Educational Research, Schloßstraße 29, 60486 Frankfurt am Main, +49 (0)69-24708-246, kuger@dipf.de

This paper was prepared for the National Academy of Education's Workshop Series on Methods and Policy Uses of International Large-Scale Assessment (ILSA). The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305U150003 to the National Academy of Education. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.Education. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education.

1. INTRODUCTION

This paper summarizes the latest practices and research topics in questionnaire use for international large-scale assessments (ILSAs). We point to the most important aspects in questionnaire design and development for international studies and highlight current challenges for the cross-cultural measurement of context factors in education. Finally, we open the discussion for research and policy issues that might lead to recommendations concerning an improved usage of context questionnaires in future studies. While we provide insight into a range of different studies, many of our examples will focus on the Programme for International Student Assessment (PISA), one of the best-known ILSAs and our area of expertise.

1.1 Current International Large-Scale Assessments

While PISA lately has almost become a synonym for ILSA of students, it was not the first of its kind. The idea of ILSA dates back to the 1950s when a framework for ILSA was developed, initiated by the International Association for the Evaluation of Educational Achievement (IEA), and implemented in 1964 by the First International Mathematics Study (FIMS). Since then, the scope of ILSA has broadened, and different organizations have taken the lead in initiating comparative studies (Heyneman & Lee, 2015). Currently, the most prominent examples of ILSAs are (in alphabetical order) the International Early Childhood Education Study (ECES), the International Early Learning Study (IELS), the Programme d’Analyse des Systèmes Educatifs de la CONFEMEN (the standing committee of ministers of education of francophone African countries; PASEC), the Programme for the International Assessment of Adult Competencies (PIAAC), the Progress in International Reading Literacy Study (PIRLS), PISA and PISA for Development (PISA-D), the Southern and Eastern Africa Consortium for Monitoring Educational Quality (SACMEQ), the Teacher Education and Development Study in Mathematics (TEDS-M), and the Trends in International Mathematics and Science Study (TIMSS) (for an overview, see Lietz, 2017).

Depending on the source of information (i.e., different participants) and its perspective on education, various angles of the educational system and its conditions, processes, and outcomes are typically assessed in addition to cognitive tests. While there have been attempts to assess instructional quality by means of international video studies (e.g., the TIMSS Video Study, and currently, the Teaching and Learning International Survey [TALIS] Video Study), most aspects can only be addressed by asking stakeholders directly. Hence, context questionnaires have become increasingly important.

1.2 The Increasing Importance of Questionnaire-Based Context Information

ILSAs in education aim to provide indicators for educational monitoring so that countries can compare their system’s performance against benchmarks, with other countries, and over time. Educational monitoring also delivers valuable information about equity and effectiveness of educational systems across all students and for particular subgroups (Klieme & Kuger, 2016). ILSAs therefore gather empirical data on student learning through establishing well-defined, measurable indicators based on educational theories. These indicators include prerequisites, processes, and outcomes of student learning, which can involve both cognitive (e.g., knowledge, competencies) and non-cognitive aspects of learning. The goal of any ILSA is to refine programs and improve student learning, which may also include competencies acquired outside of school.

Reports on ILSAs often distinguish between cognitive tests and questionnaires, the latter delivering so-called "indicators" for the learning context. Such indicators can be understood in terms of a construct to be measured. At the very least, these indicators are related to the central outcome variable of a study, but they can also be understood in the sense that they should indicate something meaningful in the educational system for the time being, for measuring change over time, or for regarding future prospects (Kaplan & Elliott, 1997).

The intention to use questionnaires to explain outcomes at the student level has always been a clear goal of ILSAs, and background data are essential to the estimation of achievement (Rutkowski & Rutkowski, 2010). In addition, context indicators as outcomes of education or as an area of interest in themselves were included in the first large-scale assessments. For example, in 1964, FIMS used questionnaires for students, teachers, school principals, and experts on education in the participating countries (Husén, 1967; Postlethwaite, 1967). Still, reporting of the results mainly focused on country rankings in students' test performance and equity measures derived from a limited number of questionnaire items.

Even though questionnaires have been included in the ILSA's design from the very beginning, the rationale behind the selection of questionnaire constructs and their importance has evolved. The third TIMSS report (i.e., TIMSS 1995) on school contexts for learning and instruction (Martin et al., 1999) does not refer to a theoretical framework for context questionnaires, but reports on the different aspects at and within the school level, with distinguishing factors including locus of control, staffing and organization, policies for classroom teaching, resources, and the atmosphere of the school. Subsequent frameworks for TIMSS 2003 and TIMSS 2007 further specified the areas of the curriculum, school, teachers, classroom activities, and students (Mullis et al., 2001, 2005). The assessment frameworks of the first PIRLS in 2001 distinguished the national, the home, and the school context, including the classroom context, as relevant factors associated with the development of reading literacy (IEA, 2000). All of these indicators can be related to students' learning while referring to different levels within the system and assuming a variety of mechanisms of educational effectiveness.

At the outset, ILSA questionnaires were referred to as "background questionnaires." Few questionnaire indicators were used in the beginning of reporting ILSA data, mainly when controlling for effects of gender or immigration background. In the 1990s, the Organisation for Economic Co-operation and Development (OECD) described the aims of PISA as to provide "contextual indicators, showing how such skills relate to important demographic, social, economic, and educational variables," and to provide "indicators on trends that will emerge from the ongoing, cyclical nature of the data collection and that will show changes in outcome levels, changes in outcome distributions, and changes in relationships between student-level and school-level background variables and outcomes over time" (OECD, 1999, p. 10). Still, a framework for questionnaires distinguishing different levels of assessment for context indicators and specifying their importance for learning was not introduced before PISA 2009 (Jude, 2016).

Only in later cycles of ILSAs did the questions of *how* and *why* students learn gain importance in study design and reporting. The majority of reports now focus on context variables at the different levels of the educational system, as well as so-called "non-cognitive" indicators. Today, ILSA instruments that gather information about the conditions and the circumstances of learning collect far more information than only the student background (Bertling, Borgonovi, & Almonte, 2016; Kuger, Klieme, Jude, & Kaplan, 2016). Reporting has increasingly addressed the questions of how and why students learn (or why they do *not* learn) and which educational provisions in a system appear to facilitate certain developments and support student learning or motivation for specific subgroups of students. Thus, each release of an ILSA is accompanied by a number of thematic reports on different topics (e.g., TIMSS Encyclopedia and PISA in Focus).

There is also a growing interest in using these indicators for repeated and trend reporting, as well as country-specific national publications (e.g., Hanushek, Peterson, & Woessmann, 2012; Huebener, Kuger, & Marcus, 2016).

1.3 Developing Questionnaire Frameworks

The more comprehensive context questionnaires become, the more important the development of a strong study framework is; specifically, one that provides (1) close links between context assessment domains and instruments, (2) thematic priorities that guide the selection of topics included in assessments, and (3) ideas on reporting.

Regarding their conceptual foundations, ILSA studies often draw on educational effectiveness research (EER). These models typically categorize relevant factors of learning into four groups of variables: context, input, process, and outcome (e.g., Purves, 1987; Teddlie & Reynolds, 2000), and stress the different levels of the educational system (Scheerens & Bosker, 1997; for a summary, see Klieme & Kuger, 2016). The advantages of basing ILSA frameworks on an EER framework are that (1) EER acknowledges the complexity of educational systems (e.g., Teddlie & Reynolds, 2000), (2) EER frameworks ultimately aim to explain student outcomes (e.g., achievement, performance, motivation, interests, engagement), and (3) overarching EER theories offer a number of different anchors to relate to other interdisciplinary theories or frameworks (e.g., Creemers & Kyriakides, 2010).

Educational effectiveness of schools and students' test performance have always been an integral part of ILSAs (for comprehensive models of school effectiveness, see Creemers & Kyriakides, 2010; Scheerens, 2015; Scheerens & Bosker, 1997). However, schooling not only aims to improve cognitive skills and accumulate knowledge, but also to foster an individual's holistic development to become an independent, active member of society. Therefore, education is also intended to support the development of non-cognitive[1] personal characteristics.

Two reasons can be given: Firstly, research has long shown correlations between student achievement and other personal characteristics such as interests, motivations, aspirations, personality traits, inclinations, values, and opinions (Hattie & Anderman, 2013). Secondly, these characteristics are related to broader life outcomes and are thus important outcomes of education in their own right (Bertling et al., 2015; Heckman, Stixrud, & Urzua, 2006; Klieme & Kuger, 2016; Rychen & Salganik, 2003; Schulz et al., 2016). Therefore, non-cognitive personal characteristics have received increasing attention in recent ILSAs (Jude, 2016; Kuger et al., 2016).

Non-cognitive constructs encompass affective, volitional, and personality characteristics, as well as attitudes and values. Several of these constructs have been incorporated into educational policies and educational system development strategies (e.g., Fadel, Bialik, & Trilling, 2015); some examples include the value of education and the motivation to engage in lifelong learning (Commission of the European Communities, 2000), problem-solving abilities and skills (Csapó & Funke, 2017; OECD, 2017a), and personality and other self-evaluative character constructs (Bertling et al., 2015), as well as a student's political and ethical mindset (Schulz et al., 2016) or well-being (Borgonovi & Pál, 2016; OECD, 2017b). Taking into account the breadth of topics considered in these frameworks and their implementation in ILSA questionnaire instruments, as well as the complexity of instrument design, Klieme and Kuger (2016), for instance, suggest using the term "context assessment" instead of "background questionnaires."

2. STATE OF THE ART IN ILSA QUESTIONNAIRE DESIGN AND DEVELOPMENT

Although ILSAs have been implemented for a number of decades, they are constantly being changed to improve their conceptual foundations, assessment methodology, implementation procedures, and analytical strategies. To adequately frame current challenges of and recommendations for ILSA, the following section highlights selected aspects of state-of-the-art questionnaire design in ILSAs. We will discuss the relevance of sound theoretical frameworks for questionnaire development, the importance of balancing trend measures and new content, the latest developments in computer-based assessment, and international translation and verification procedures.

2.1 Conceptual Foundations: Frameworks and Links to Theory

Depending on the intended policy targets in ILSAs, frameworks can take into account different aspects, although not all frameworks are detailed enough to cover all aspects. Such aspects include (1) the definition of the construct(s) to be measured, that is, whether the construct be defined and adequately operationalized in all participat-

[1] The research literature often differentiates between cognitive and non-cognitive outcomes (Gutman & Schoon, 2013; Heckman, Stixrud, & Urzua, 2006). The label "non-cognitive" is used in full awareness that many of these characteristics, values, or attitudes rely on cognitive processes. The term is rather used to clearly distinguish measures of cognitive capability or performance from the broad range of other personal characteristics (for a discussion of this terminology, see Humphrey et al., 2011).

ing countries; (2) a specification of the measurement approach, that is, which format the instruments should employ; and (3) a specification of the method of analysis to determine the extent of comparability of the results, that is, which statistical procedures and scaling models will be used to establish comparability and what would be done if comparability is not fully supported (see section on invariance). Thus, since current ILSAs follow different goals, their frameworks differ remarkably. Although most frameworks refer (at least implicitly) to the overall notion of educational effectiveness, the studies emphasize different aspects (e.g., teaching and learning in the classroom, schools as educational settings, diverse student outcomes) based on their respective policy focus (van de Vijver, Jude, & Kuger, in review).

In ILSA questionnaires, the theoretical concepts or constructs can be distinguished according to latent variables (to be measured indirectly, i.e., motivation), manifest indicators (that can be measured directly, i.e., gender), and so-called "derived variables" that use a combination of indicators (manifest and/or latent, i.e., students' socioeconomic backgrounds). Frameworks typically include a theoretical rationale on why each construct is important for reporting and how it should best be assessed; one example is by asking specific stakeholders in education, such as parents or teachers, rather than students.

Frameworks also need to take into account the assessment design with respect to the amount of time assigned to the context questionnaires, that is, the time that students, schools, and other stakeholders will be expected to spend on responding to a survey. This allocated time per questionnaire, but also per construct or area to be covered, presents a baseline for the amount of material that can be included in the assessment.

Given that ILSAs aim at measuring change over time, not only regarding outcomes but also reflecting on context measures, frameworks need to set priorities in content selection to find a balance between trend measures and innovative constructs.

2.2 Balancing Trend and Innovation in Questionnaire Development

The majority of ILSAs are aimed at trend reporting, that is, describing change over time for specific indicators. Questionnaire development thus substantially relies on screening instruments from previous cycles and selecting content that was chosen for trend reporting. Strictly speaking, trends can only be reported if the exact same instruments are reapplied; as stated in a quote attributed to Samuel Messick: "If you want to measure change, don't change the measure" (National Research Council, 1999, p. 77). Obviously, cultures and educational systems differ, and even within a singular system, learning cultures can change. Thus, the meaning of certain measures may not remain stable over time. A good example is the question assessing household resources that was updated to now include smartphones or tablets. Considering the switch to electronic readers and online reading, one could also question why the "number of books per household" is still being used as a prominent indicator in TIMSS and PIRLS. Consequentially, small adaptations are sometimes necessary to reapply instruments from previous cycles. Although the results need to be inspected with caution, reporting such conceptual trends may be more ecologically valid and therefore more meaningful than applying outdated measures that fulfill full technical trend criteria.

Another important consideration is the degree to which an ILSA should relate to similar studies. Regarding certain aspects of reporting, it might, for example, be desirable or advisable to link the U.S. National Assessment of Educational Progress (NAEP) to an international assessment such as TIMSS or PISA. Given certain preconditions, this could be done by using an overlapping sample or by applying the same assessment instruments and questionnaires in both studies. A combination of the strengths of two studies by establishing such a link can lead to additional insights or a more refined power of analysis (e.g., Kaplan & McCarty, 2013).

In addition, a broader scope of the context-assessment instruments needs to be taken into account, including, for example, newly stressed reporting categories (e.g., more information about immigrant background, well-being, or student non-cognitive characteristics) or overall changes in survey methodology (e.g., increased attention to equivalence in measurement quality, more detailed targeting of questions to certain subgroups of students). Furthermore, changes in ILSAs might become necessary because of the introduction of new item formats. Likert Scale–type question formats, the most commonly used formats in questionnaire assessments, have long been challenged by research findings, particularly in international comparisons where they are often implemented with a comparatively small number of response categories (Hui & Triandis, 1989). One particular weakness of Likert Scale–type items is their tendency to reflect different kinds of response bias; for example, the phenomenon of some cultures choosing the more extreme answers (e.g., "strongly agree") over the moderate ones (e.g., "somewhat agree") (Hartley & MacLean, 2006; Lee et al., 2002). Thus, recent surveys have more frequently introduced innovative formats and methodological alterations to reduce problems associated with 4- or 5-point categorical answering scales (e.g., OECD, 2014). Among the many different possibilities are, for example, forced-choice item formats (Bartram, 2007; Xiao, Liu, & Li, 2017), situational judgment questions (O'Connell et al., 2007; Patterson et al., 2012), the over-claiming technique (Paulhus et al., 2003), and anchoring vignette techniques (Hopkins & King, 2010; King & Wand, 2007). Evidence exists that some of these measures help to decrease answering biases (Kyllonen & Bertling, 2013; OECD, 2014). Still, there is a clear need to gain further knowledge on how and to what degree new item formats facilitate data collection and processing and increase data quality and comparability while allowing for trend comparison across assessment cycles (Lietz, 2017; Rutkowski & Rutkowski, 2010).

2.3 Computer-Based Assessment

One important innovation made for many recent ILSAs is the introduction of new item formats as a result of technological improvements: PISA transferred to computer-based assessment for its 2015 cycle, and TIMSS 2019 will be administered electronically.[2] Such recent developments are accompanied by certain advantages to reduce the impact of previous shortfalls. Computer-based assessments deliver formats that facilitate question comprehension, question formatting, and the answering process itself. Indeed, the interactive nature of digital assessments can illustrate a participant's answer or provide feedback. Questions asking for percentages or amounts can be illustrated by interactive

[2] The assessments ePIRLS and eTIMSS are by definition online assessments and are thus designed to work on digital platforms. A transition from paper and pencil to online formats was not necessary and problems can be assumed for assessments that are subject to a mode change (see Buerger et al., 2016).

graphs, for instance, illustrating percentages in a pie chart while the participant enters numbers or amounts. In sorting tasks, presenting the final order of stimuli (as sorted by the user) provides a higher degree of clarity than a list of rank orders next to the stimuli in their original sorting. The answering process can also be facilitated by providing a more streamlined layout of a question in an assessment (e.g., replacing long lists of answering alternatives with drop-down menus). Furthermore, computer-based assessment platforms can help reduce answering biases. For example, video vignettes of classrooms with different degrees of disciplined behavior may be used to anchor participants' interpretation of a question or additional background information can be added via pop-up help sections (e.g., definitions of moderate and vigorous physical activity, additional answering instructions; Kunz & Fuchs, 2018).

Finally, computer-based assessments can improve data quality and facilitate data post-processing. Reminders to answer questions that the participant has skipped can help reduce the amount of missing data, and plausibility checks can be introduced in ongoing assessments. Both can be applied to prompt participants to reconsider congruency of different answers (e.g., "100 hours" each of mathematics, science, and reading lessons per week in an overall schedule of 32 lessons per week), to answer previously unanswered questions, or to correct implausible or unlikely answers (e.g., a sum of more than 120 days of deviation from the regular school schedule due to weather phenomena, teacher strikes, and schoolwide activities). Although such reminders seem to be an attractive solution to common problems in questionnaires, they should be implemented with care. High pressure to answer all of the questions may cause reactance and lead to higher drop-out rates (Décieux, Mergener, Neufang, & Sischka, 2015).

2.4 Translation, Adaptation, and Verification in Questionnaire Development

ILSAs are aimed at comparing contexts, processes, and outcomes of learning. A prerequisite for comparison is that comparable constructs are measured in different assessment contexts, such as different languages and cultural settings (Harkness et al., 2010b). There is an important difference regarding the translation of cognitive test and questionnaire material. While the translation of cognitive test material must take care not to change the difficulty of a test question or answering option, the translation of questionnaire material must ensure that the meaning of a question is kept in the respective cultural context. Simple answers such as "yes" or "no" can easily be assumed to be comparable in different cultures and languages, but established rating scales using categories such as "strongly agree" or "somewhat agree" might need to be treated with more caution when it comes to the comparison of different grades of agreement across languages (Dept, Ferrari, & Halleux, 2017).

A broad area of research is dedicated to the translation of international surveys, and consequently, recent ILSAs have reflected state-of-the-art approaches, such as double translation (i.e., when the international master versions are prepared in two different languages and national versions are thus translated twice and compared) instead of translating only from one language to another (Grisay et al., 2007). In addition, newly developed questionnaire material is tested in cognitive laboratories in different countries to understand how test-takers interpret the material. ILSAs have most recently incorporated so-called "translatability assessments" into the process of developing new questionnaire material. Translatability assessment is innovative in terms of translating new questionnaire content into several language groups (e.g., Arabic, Korean, Roman, or Slavic) during the development phase in order to discover any difficulties that the content or the wording might pose for translations and the comparability of meaning (Dept, Ferrari, & Wyrinen, 2010).

In ILSAs, test material is usually kept unchanged across all language groups, but questionnaire material has to be adapted to fit the national context. For example, study programs (e.g., academic and vocational tracks and streams) or educational levels (e.g., primary, lower secondary, upper secondary) most certainly will be country-specific. Indicators of wealth or home possessions might also deviate from country to country, such as "antique furniture" in Italy or "a home cinema" in Korea (OECD, 2016b). For this reason, a sophisticated process of so-called "national adaptation" is included in all ILSAs, assuring that countries choose national indicators that can be compared on an international level, for example, by aligning educational levels to the International Standard Classification of Education (ISCED) levels. Such "functional equivalence" assumes that questions may differ in content, but ask about the same construct or phenomenon as the meaning is kept and adapted to the national context (Harkness et al., 2010a). In ILSAs, translation and adaptation of questionnaire material is a multi-step process that involves several stakeholders, such as linguistic and assessment experts, to ensure translation quality as a crucial requirement to guarantee data quality. Another requirement for those experts is their in-depth knowledge of cultural particularities in their respective educational system (Ebbs & Djekić, 2016; OECD, 2017).

2.5 Questionnaire Design, Content Coverage, and Assessment Time

Questionnaires are an essential part of ILSAs that have to be integrated with an overall assessment design. This design defines the order of administration (e.g., whether questionnaires follow the cognitive assessment), as well as the length of each assessment instrument. So far, studies have varied only slightly regarding the design of questionnaire administration: TIMSS 2015 implemented a 30-minute student questionnaire after the cognitive assessment (Mullis & Martin, 2013), a so-called "home questionnaire" for the parents of the fourth-graders' sample, a 30-minute teacher questionnaire for the mathematics and science teachers of the students sampled, and a 30-minute questionnaire for the principals of each participating school. The questionnaire design for PISA 2015 was similar, also including a teacher, principal, student, and parent questionnaire component. Students were granted up to 35 minutes to answer the overall questionnaire and two so-called "international options" were offered, the latter to be chosen on a country level. These additional questionnaires focused on aspects of the educational career (EC) and information and communications technology (ICT) familiarity, with each one adding 10 minutes of assessment time to the overall design for the students in the participating countries.

This solution of offering optional add-ons to the common student questionnaire shows that the increasing interest in assessing context factors creates the need to either extend the overall assessment time for questionnaires or select content for each new assessment cycle and thereby run the risk of losing trend information on key indicators. In the section on conceptual frameworks, we stress the fact that the study framework should provide a sound rationale for the inclusion of topics in the assessment. Yet, for studies involving a heterogeneous set of countries, it might be feasible to declare a certain share of questionnaire content to be common content (i.e., mandatory for all countries and probably including trend material; see section on balancing trends and innovation in questionnaire development). Moreover, a degree of freedom would be granted for countries to choose from a prepared set of options that are all integrated within an overarching framework but are perhaps of less interest for some countries (e.g., background information from students with immigrant status). Again, selection should be made based on the assessment framework, yet different design options could be discussed to broaden the scope of questionnaire content over time.

One approach to increasing the content coverage of topics of interest without increasing individual students' response time was implemented in PISA 2012, which is, to our knowledge, the only ILSA with a booklet-rotation design for the questionnaire content (OECD, 2014). Three overlapping questionnaire booklets were administered, allowing for the assessment of a broader range of constructs, albeit creating a missing data pattern for constructs that were not administered to the same students. While the estimation of the cognitive measure was not influenced by the booklet design (Adams, Lietz, & Berezner, 2013; Kaplan & Su, 2016), the potential for enhanced analysis was limited. The problem with such an approach is that constructs administered in different booklets are answered by different students and thus cannot be related to each other. Consequently, one of the many challenges of ILSAs concerns the enhancement of the explanatory value by broadening questionnaires, as described in the next section.

3. CURRENT CHALLENGES

Some researchers have pointed out that context questionnaires in ILSAs currently face a great number of challenges. Although many improvements have been made, there is still ample room for improvement in many aspects (e.g., Rutkowski & Rutkowski, 2010). In the following sections, we highlight some of the more pressing challenges that need to be addressed, including the measurement of students' background, the importance of developing new outcome measures, the potential of computer-based assessment, and challenges for future assessment design, as well as the issues of data quality and measurement invariance of context factors in international comparisons.

3.1 Measuring Socioeconomic Background and Diversity

Equity in education presents a major topic for educational monitoring (Klieme & Kuger, 2016). Equity can be defined according to different aspects, but it is usually reflected in terms of the socioeconomic status (SES), immigration status, and gender as common equity criteria across countries. The importance of SES for educational achievement has frequently been shown, yet the indicators or measuring approaches have varied significantly, resulting in different correlations between SES and achievement (Buchmann, 2002; White, 1982). The most important indicators are parental education, parental occupation, and family income, but ILSAs vary in how they operationalize the assessment of background variables in their questionnaires (for an overview, see Brese & Mirazchiyski, 2013; Watermann, Maaz, Bayer, & Roczen, 2016).

Several measures that aim to reflect these indicators exist in ILSAs and can be obtained from questionnaire respondents, such as the highest index of SES for which the higher value of the parents' occupation is used, the highest educational attainment expressed in years of education, indicators for home possessions, the Erikson-Goldthorpe-Portocarero (EGP) class scheme (Erikson et al., 1979), and differentiating between different values of employment status and type of work, to name a few (see Leiulfsrud, Bison, & Jensberg, 2005). Indicators of social background in ILSAs underwent substantial development, first assessing mainly the father's education and occupational status (for an overview, see Buchmann, 2002), but later also asking about social and educational resources at home, and finally developing composite indicators as used in PISA (Watermann et al., 2016).

Currently, the PISA study by OECD uses the index of economic, social, and cultural status (ESCS). ESCS is a composite index consisting of three factors, including parents' occupation, their education and vocational training, and home possessions. This measure has been criticized regarding its reliability and cross-cultural comparability (Ortmanns & Schneider, 2015; Rutkowski & Rutkowski, 2013).

Furthermore, the validity of such measures can be questioned when students are asked about their parents' education and occupation, but pertinent research suggests that the results can be considered valid (Jerrim & Micklewright, 2014). According to Sirin (2005), the correlation between SES and student achievement is stronger when seeking information from parents themselves rather than asking the students, yet more missing data are usually found in the parent questionnaires. In current ILSAs, parents are either asked to classify their profession into different categories, ranging from small business owner to professional (as in TIMSS and PIRLS), or open-ended questions in the student questionnaires are used to assess parental occupation. In PISA, for example, students' answers are coded into numeric classification codes called ISCO codes that should be internationally comparable, having been developed by the International Labour Office (International Labour Office, 2012). Manual coding is rather expensive and might be prone to errors or missing information. For PISA 2015, up to 7 percent of the data was missing for at least one of the indicators composing the ESCS index. To improve data quality, computer-assisted measurement and coding has been suggested, and development projects are underway (Gweon et al., 2017; SERISS, 2017; Tijdens, 2015).

The analysis of equity or SES as a moderating factor is even more challenging when it comes to comparisons with low-income countries, not only in regard to enhancing the measures that are implemented, but also the models for analyzing relationships. Willms and Tramonte (2014) discuss different approaches of PISA for development to modeling nonlinear relationships between educational achievement and SES or identifying thresholds for maximum impact of SES in specific countries. They also suggest the development of "poverty measures" in addition to measures of wealth.

In the most recent round of PISA, measures of students' socioeconomic backgrounds in addition to wealth were implemented with the Field Trial, aiming at a more differentiated approach to assessing students' socioeconomic background (Waterman et al., 2016). These measures included questions on families' cultural activities, social support, and beliefs, but also school indicators such as free lunch programs and subsidized textbooks or school trips. TIMSS asked the principal about the percentages of students in the school coming from economically disadvantaged homes. Future ILSAs might take into account that different measures of SES might lead to different conclusions, and that some measures might be more appropriate to explain variance for specific groups of students or when comparing data among groups of countries.

3.2 Outcome Measures: Inconsistency in Inclusion and Measures of Non-Cognitive Outcomes

Research evidence strongly and consistently points to the importance of non-cognitive skills that are predictive of achievement outcomes (e.g., grade point average and test performance or attainment), later labor market success, and general life outcomes (e.g., Almlund, Duckworth, & Heckman, 2011; Heckman & Kautz, 2013; Heckman et al., 2006; Lindqvist & Vestman, 2011; Richardson, Abraham, & Bond, 2012). Simultaneously, but hardly independently, non-cognitive outcomes have played an increasingly important role in policy making and education administration. General frameworks of education for future generations incorporate attitudes, values, and character non-cognitive outcomes (Tawil, 2013; Trilling & Fadel, 2009). Consequently, curriculum changes in many countries now emphasize that education and schooling should acknowledge these outcomes as important results of education next to academic achievement.

To adequately reflect this multidimensionality of education goals and to report on non-cognitive outcomes as well as such precursors of student achievement, there seems to be a wide consensus that ILSAs should not only capture test performance and/or attainment but also other results of education (e.g., Bertling et al., 2016; Klieme & Kuger, 2016). Still, there is unfortunately little agreement on which aspects qualify for inclusion, which are most important for policy decisions, and how an overall framework could incorporate the different aspects.

Frameworks that are grounded in economics and psychological research often stress the importance of personality traits (Bertling et al., 2016; Brunello & Schlotter, 2011; Heckman et al., 2006; John & De Fruyt, 2015). The “21st century skills” approach encompasses an even broader range of constructs (Trilling & Fadel, 2009): Fadel and colleagues emphasize agency, attitudes, behaviors, dispositions, mind-sets, personality, temperament, values, beliefs, social and emotional skills, non-cognitive skills, and soft skills (Fadel et al., 2015, p. 127). Additionally, aspects of well-being and quality of life are often implemented in surveys on health and development (Bertling et al., 2016).

All of the constructs mentioned above are without a doubt important outcomes of education. Likewise, frameworks, policy measures, and educational processes at all levels of the educational system—that is, meta-governmental frameworks, national curricula, syllabi, school and classroom provisions, and teaching processes—need to focus on fostering their development in students’ lives. Unfortunately for ILSAs, existing disagreements cannot simply be resolved. There is too little research consensus on the number of internationally equivalent and relevant outcomes to be included in ILSAs, adequate theoretical frameworks to organize non-cognitive outcomes as a group, and too little research on their combined interplay with educational achievement. Important tasks for researchers will be to better define the uniqueness of each construct or overlaps of differently labelled but similar constructs and to describe educational practices and settings to support students in developing cognitive and non-cognitive outcomes equally well.

3.3 Computer-Based Instruments and Their Potential Power for Context Assessment

While many ILSAs are still paper-based, assessments like PIAAC, PISA, and NAEP have made the transition to computer-based formats. Electronic assessment not only facilitates operational procedures and improves data quality, but it also has the potential for additional analyses based on so-called "log data."

Log data are electronically recorded during the process of responding to the computer- based assessment. These log data pertain to the content of answers and the time needed for answering questions, but also to the navigation among different pages of the electronic assessment. These data can be used for different purposes, including measuring the time it takes to work on tasks (which can be an indicator of accuracy and can also inform on students' test-taking behaviors [OECD, 2015]) or more specific aspects such as reading capacity and engagement in addition to the test score obtained (Naumann, 2015). The PIAAC study already published an online LogData analyzer tool that allows for easy access to the data for secondary analysis.

Log-file data in educational research have been used to relate answering behavior to cognitive processes (Almond et al., 2012) and to identify reoccurring sequential action patterns that can be associated with success or failure in tasks (He & Davier, 2016). Regarding the measurement of context factors, log-file analysis has been used to measure motivation, a factor that is traditionally assessed by asking students self-report questions (Hershkovitz & Nachmias, 2009). "Big data" approaches are already being used in psychological research, for example, by linking answering behavior in computer-based testing to personality aspects that could explain learning behavior (Papamitsiou et al., 2017). Other publications focus on the usage of data mining and learning analytics to analyze students' learning styles in electronic learning and testing settings (Agudo-Peregrina et al., 2013; Efrati et al., 2014).

As this area of research is still quite young, to our knowledge, no theoretical frameworks exist specifying which kind of log-file data would be the most promising to contribute additional information in ILSAs. Notably, research is still missing on the relationship between context indicators as assessed by tests and questionnaires and corresponding data from log-files. Moreover, no educational theories come to mind that would link answering behavior to questions about learning contexts. Accordingly, a debate has evolved about the usefulness of big data in educational contexts and educational data mining (Cope & Kalantzis, 2016), with big data being defined as "the purposeful or incidental recording … of varied types of data" (p. 2), and how to balance the benefits and the risks of aspects such as data privacy (NAEd, 2017; Williamson, 2017).

There is no doubt that electronic data can help visualize learning processes or predict learning outcomes by easily collecting and recording large datasets (for an overview, see Romero & Ventura, 2010). However, sound theories of relationships between these context factors of learning processes and learning outcomes are still scarce, and current approaches remain exploratory. Future model-based approaches would need to identify causal relationships between the available indicators and the learning outcome (Harlow & Oswald, 2016). The accessibility and the durability of these data need to be examined along with methods of data analytics and usability in learning and assessment. For example, the European Union currently supports research in this area (EU, 2017), funding projects using PISA log-file data to specify how these data can be used to "contribute to the assessment of non-traditional skills and competences" (p. 9), some of which are currently being measured with context questionnaires. One question that seems to be completely undiscussed is how sensitive big data are to country-specific variation due to country-specific electronic infrastructure.

3.4 The Issue of Measurement Invariance

Data from context assessments are prone to a number of biases. Failure of invariance of measurement across countries, therefore, is an important aspect that has received increasing attention in recent years in the context of ILSAs (van de Vijver & He, 2016; van de Vijver & Leung, 1997). Equivalence of measurement of a construct in different subgroups is a prerequisite for comparisons. Researchers have commonly discussed three levels of invariance that should always be tested and confirmed before using a construct for country comparisons (see Nagengast et al., 2014, for more details and an empirical example of applying more nuanced levels on PISA data). For one, invariance on the configural level confirms that a construct exists in all participating countries and that the chosen set of items can be used to measure it (i.e., all items load on the latent factor in all countries). Secondly, metric invariance is confirmed if all items contribute equally to the measured construct (i.e., equivalent item loadings on the latent factor in all countries); such constructs can be used in parallel correlation analyses in all countries (e.g., correlating a certain teaching indicator with student achievement in each country). Thirdly, to compare construct means across countries (e.g., students in one country are more interested in a certain topic than students in another country), scalar invariance must be confirmed for the construct under study. In technical terms, scalar invariance holds if all items have identical intercepts in all countries.

Examples from PIAAC (Gorges et al., 2017) show that context scales can achieve invariance across gender, age group, level of education, and immigrant background in different language versions and can therefore be used to compare means across participating countries. However, TEDS-M, which focuses on an adult population, shows that non-cognitive indicators may vary by country but can sometimes be compared among selected cultures for which scales prove to be invariant (Laschke & Blömeke, 2016). Invariance testing is therefore essential before attempting to compare questionnaire scales among countries or groups within a country.

There are different reasons for why scalar invariance is rarely achieved (e.g., Rutkowski & Svetina, 2017) in ILSA context assessment data, and the literature proposes different ways of dealing with this issue. Most approaches either focus on questionnaire design and item format (i.e., measures implemented before data collection) or they apply post–data collection modeling techniques. According to one proposal, low invariance across countries in questionnaire design can be dealt with by the use of certain item formats (e.g., video vignettes, situational judgment methods) that are supposed to prime participants in a particular manner and thus reduce response styles or differential construct interpretation (Kyllonen & Bertling, 2013; see section on balancing trend and innovation in questionnaire development). Lower levels of measurement invariance can also be accounted for by allowing for approximate invariance (Asparouhov & Muthén, 2014) or for partial invariance (Rutkowski & Rutkowski, 2017) during scaling or secondary data analyses.

The approach of introducing anchoring vignettes (King & Wand, 2007) requires the inclusion of an extra question in the context assessment—not just an alternative format for a given question—and particular data post-processing. Individual response styles are captured in a participant's answer to certain vignettes, and a self-report scale can then be corrected according to each individual's answer to the vignettes. Recent research has shown that such a procedure can increase a scale's validity (He, Buchholz, & Klieme, 2017) but not necessarily its reliability (Marksteiner, Kuger, & Klieme, in review). In addition to methodological approaches, qualitative aspects can be used when explaining lack of invariance, such as cognitive interviews (Collins, 2003). Although different response styles and biases are mainly being discussed in the context of cross-country comparisons (van de Vijver & Leung, 1997), there is evidence of differential answering behaviors within countries and languages as well (Eigenhuis, Kamphuis, & Noordhof, 2017; Yap et al., 2014).

Hopfenbeck and Maul (2011) investigated another aspect of comparability, showing a link between students' science literacy and the amount of valid answers in learning strategy scales, highlighting a bias for low-achieving students. They suggest introducing different answering scales and reducing the reading load to reduce bias in context measures depending on the cognitive competence.

3.5 Non-Response in Questionnaire Data

Non-response to questionnaires has become a heavily debated topic (Mohadjer et al., 2013; OECD, 2010; Rutkowski & Rutkowski, 2010), as failure to respond might be non-random, that is, influenced by factors such as language competence, educational level of the responders, school characteristics, or the construct to be measured. Even though countries need to reach a certain sampling threshold to be fully included in the international reports, bias needs to be estimated to evaluate its impact on results and conclusions from the analysis. Bias can be analyzed by using national or school indicators to compare characteristics of non-respondents with those who do respond (e.g., language background, socioeconomic indicators) (Blom, Jäckle, & Lynn, 2010).

Data from PISA, for example, show that parent questionnaires are more frequently returned by parents who speak the test language and have a higher SES. Accordingly, interpretation of the results needs to take this limitation into account. Meinck, Cortes, and Tieck (2017) give an overview of the scope of non-response in IEA studies in the past 10 years. They conclude that "IEA studies face a non-negligible amount of non-response, which occurs especially at [the] school level in student surveys and at both sampling stages when adults are the target population" (p. 6). Thus, if schools decline participation, or if sampled principals, teachers, or parents do not respond, bias analysis is the method of choice to document the degree of biases included in the data. The authors' proposal to implement shortened, non-respondent questionnaires might be a possible remedy to reduce the risk of bias in ILSAs. Non-response might be motivated by the topics addressed in the questionnaires, but also by the sheer length of the assessment time.

4. RECOMMENDATIONS FOR SYSTEM MONITORING POLICIES

It has been argued that the goal of large-scale assessments—that is, mainly the comparison of educational systems and their outcomes—might not directly lead to policy-relevant results. Different reasons can be discussed. Rutkowski and Delandshere (2016) present different arguments for treating these data cautiously. Their main point is that ILSA indicator systems can neither reflect the full complexity of educational systems nor the dynamic interactions of context factors over time. They also point to the question of validity in questionnaire scales depending on the selection of variables representing a specific construct and the operationalization of the outcomes in different ILSA studies. Consequently, certain conditions must be given when addressing causal questions within ILSA contexts and, as Kaplan (2016) argues, the current designs do not allow for drawing causal inferences from ILSA data. The direction of causality is not always clear, and confounding variables that were not measured might explain existing relationships (Allen & van der Velden, 2014). In the following sections, we thus present some recommendations based on current practices and challenges mentioned in this paper. Several aspects are highlighted that should trigger a broader discussion between educational researchers and policy makers on how to overcome the existing shortfalls in context measurement.

4.1 Interplay of ILSAs and Educational Research

The identified challenges and the current state-of-art approaches to context assessment in ILSAs lead to several different recommendations for researchers as to how their work could substantially help broaden the scope of ILSAs. The following paragraphs highlight three such recommendations that could be implemented with rather little effort and have seldom been discussed elsewhere. They cover different aspects of framework improvement and construct validity, as well as closer links between and better coordination of national and international research on ILSA data.

The main goals of ILSAs are derived from policy questions, which guide or back up educational policy making. Thus, the driving force for ILSA development and implementation is typically societal interest in certain aspects of education and respective decision making (e.g., equity or effectiveness). By contrast, disciplinary research in education-related disciplines is usually guided by constructs and relating theories (of state and change). ILSA frameworks would profit from more research involvement in multiple ways: for one, construct selection during framework development would profit from more in-depth cross-country research on constructs, that is, the verification that important constructs for education reporting in one country are important in all countries regardless of cultural values, the histories of educational systems, and teaching traditions. For an ambivalent example, see Caro, Lenkeit, and Kyriakides (2016), who studied the comparability of certain teaching practices' relative importance in different countries without testing for measurement equivalence. Furthermore, more national research is needed within countries that takes up constructs included in ILSAs and compares its conceptualization and measurement approach to national research and practice (e.g., Schmidt, Burroughs, Zoido, & Houang, 2015; Zuzovsky, 2013). Such national research involvement could also contribute to finding reasons for low measurement invariance.

Moreover, context assessment in ILSAs would profit from more national and international re-analyses of the data. Marksteiner and Kuger (2016), for example, studied the dimensionality of "sense of belonging" as it was implemented in PISA 2012 and possible mechanisms of parental influence. The study revealed some weaknesses in the construct's validity regarding its dimensionality and measurement equivalence. Further analyses should take a closer look at national realizations and representations of constructs implemented internationally and thus deliver stronger proof of construct validity and policy relevance. These links between national research traditions and ILSAs should be related to the literature body on national particularities in order to better inform policy making in all countries.

To feed back resulting findings from national, cross-national, or international studies to future cycles of ILSAs, the findings need to be collected, evaluated, and systemized by research repositories. Adequate literature (or systematic) reviews need a substantial amount of time and participation of researchers worldwide, but they would help improve ILSA frameworks on the one hand and policy making on the other hand. Only if we are able to prove the importance of constructs in all participating countries and understand their mechanisms of influence and change can we use ILSAs as an equally strong tool in all countries to improve educational systems and processes.

4.2 ILSAs for Monitoring System Trends

The question of validity of data derived from ILSAs for policy decisions is not new. In addition to the comparison of outcomes, results over time can be used to compare the effectiveness of educational systems regarding achievement (see Lenkeit & Caro, 2014). Still, we argue that with such a strong emphasis on country rankings in achievement outcomes, the broader view of a monitoring function of ILSAs is lost—particularly now with a focus on non-cognitive outcomes measured in the questionnaire. To use the full potential of ILSA data, an in-depth inspection of the existing range of indicators would be advised.

When taking on a long-term perspective, change over time in important educational outcomes needs to be monitored, not only change regarding cognitive achievement. These data are, for example, available for the OECD's Indicators of National Education Systems (INES) project; see the publication *Education at a Glance* (OECD, 2017c). An important area of research concerns the development of analytical tools designed to exploit the features of indicators that ILSAs were designed for—namely in detecting changes in educational systems when such changes actually occur. This feature of indicators as tools for detecting and monitoring change was aptly summarized by De Neufville (1978): "Theoretical indicators can also be validated by looking at the movements in relation to indicators of other variables when an interrelationship is presumed" (p. 177) (see also Kaplan & Elliott, 1997; Kaplan & Kreisman, 2000).

ILSA data have been available since the 1960s for some countries, even though studies, indicators, and populations might have changed. Some countries even implement different ILSAs simultaneously, such as TIMSS and PISA, focusing on somewhat comparable aspects of the educational system. As the purpose of these studies—besides comparing learning outcomes across countries—is to observe changes in educational systems, results from these studies should be combined to address aspects of educational contexts from different angles. Moreover, combining results can also yield some kind of validation. If specific context indicators have been shown to be important for learning outcomes, can these results be replicated over time and across studies? An approach that combines or validates knowledge from different ILSAs is particularly important for policy makers, as well as educational practitioners looking for alterable impact factors. Still, differences in the design and the goals of these studies need to be evaluated before results can be compared.

While TIMSS, for example, follows a curricular approach, thus resulting in a grade-based sample and tests for mathematics and science every 4 years, PISA implements a competence-based approach, assessing reading, mathematical, and science literacy every 3 years in an age-based sample. Even though the TIMSS sample of eighth grade students will overlap with the PISA sample of 15-year-old students in many countries, the content approaches to assess learning contexts differ to a certain degree when it comes to comparing classroom practices.

Klieme (2016) highlights the following five areas of difference between PISA and TIMSS: (1) The *curriculum approach of TIMSS* that is supposed to be valid across countries, and thus includes indicators focusing more on knowledge, versus the *life-skill approach of PISA* that is reflected in the indicators being embedded in real-world problems; (2) the grade-based selection of whole classes (TIMSS) versus the age-based selection of 15-year-old students in schools and different grade levels (PISA); (3) the participation of different OECD and non-OECD countries in both studies; (4) the mode of assessment on paper (TIMSS) and the computer (PISA); and (5) the scaling approaches for both cognitive and questionnaire data that differ between the studies.

Overall, a prerequisite for harmonizing data across studies is that samples and indicators are comparable to a certain degree. To analyze this comparability, meta-data on samples, indicators, and methods of analysis need to be documented and made available. To date, there is no comprehensive database on the indicators used in TIMSS or PISA across time, that is, documentation on questions and their specific wording (including changes made between assessment cycles), the identification of comparable questions in the respective datasets for each study, and the measurement quality of scales per assessment time, among others. Additionally, indicators that exist in different studies should then be matched to facilitate comparability of measurement approaches and results across studies.

Even though considerable effort has been made to facilitate access to comprehensive databases, such as through the IEA's Data Repository and IDB Analyzer, a country-specific documentation of context indicators, national adaptations, and changes over time is still missing. While questionnaires and codebooks are available freely or by request for research purposes (e.g., from IEA [http://www.iea.nl/other-iea-studies] and OECD), no overall documentation of all context indicators per cycle is currently available. Only such documentation would allow for an easy detection of gaps in indicator systems of specific studies, as not all indicators that might be interesting for trend analysis have been implemented in all cycles of PISA or TIMSS. Owing to a lack of such meta-data in comprehensive documentation, harmonization of datasets needs to be done for each specific research purpose. Still, different procedures can be applied for ex post facto harmonization of survey data (Granda et al., 2010). In times of declining resources, this seems particularly relevant if policy decisions need to be made as to which studies to implement and which indicators to choose to evaluate the educational system in the long run.

Furthermore, no comprehensive documentation of material from field trials of these studies has been made publicly available. Developers thus run the risk of reinventing content that has been tested in previous cycles but was not taken up for international comparison because of quality issues. One exception is the latest documentation of PISA Field Trial questionnaires, which include meta-data, that have been made publicly available (Kuger et al., 2016; accessible at http://daqs.fachportal-paedagogik.de/search/show/survey/177). There are, however, national endeavors to document ILSA indicators and constructs in an easy-to-use tool. Investments in this area would soon pay off for any future development of ILSA studies, as well as harmonization approaches.

5. CONCLUSIONS—ILSA DATA FOR EDUCATION CHANGE

In recent years, several papers have been published that analyze the impact of ILSAs on educational policy in the participating countries (for an overview, see Heyneman & Lee, 2014). While there is no system of tracking or evaluating the impact directly, many policies relate to the outcomes of ILSAs, or at least claim to do so (Breakspear, 2014). However, it is important to discuss which data are used to change policies. Most reforms seem to focus on changes in curriculum, the introduction of standard-based testing, changing policies for minority students or decentralization, and funding without taking mechanisms and effects on actual practices and outcomes into account (Martens, Niemann, & Teltemann, 2016).

This opens up a debate about which data could be used, and more precisely, which data beyond mere achievement data, such as results from context questionnaires, should be taken into account when using ILSA data for changes in educational policy, schools, or even teaching practices. Accordingly, Teltemann & Klieme (2016) note that there is currently no information on how such policies affect the classroom level—for which indicators are available in most ILSA questionnaires—or on cognitive and non-cognitive educational outcomes.

When taking a look at classroom practices as assessed in OECD studies, PISA and TALIS seem to deliver interesting information from different perspectives. Even though TALIS coincides with PISA for the first time in 2018, overall samples are not linked systematically and different questionnaires for teachers and principals have been implemented. For 2012, however, a subset of countries implemented a link among the teachers working in schools sampled for PISA and provided TALIS questionnaires for teachers and school principals (OECD, 2016a). Although this approach seems to allow for comparisons of teaching and learning practices on a school level, several prerequisites need to be discussed. The sampling of both studies does not allow for a direct link between teachers and individual PISA students, beyond the fact that both are linked to the same school. Still, the report matches students' performance on PISA to indicators on teaching strategies. In our opinion, however, the relevance of analysis based on these data should be carefully considered since the age-based sample of PISA students cannot be directly linked to teaching strategies that usually refer to a classroom context or at least a certain grade level. Moreover, teacher characteristics and teaching strategies vary greatly within schools; matching student outcome measures to teacher reports that are not directly linked might go beyond what the data can reliably deliver. It seems likely that TALIS (which is historically more focused on teachers and principals) and PISA (which is historically more focused on students and principals) will be integrated and involve the same schools across countries by 2024. However, when classroom-based practices are analyzed, it remains to be discussed if an age-based sample such as PISA is sufficient to derive conclusions for teaching.

There are some indications that data on curriculum, standard-based testing, evaluation, and resources, especially from the school questionnaires, are being used in national policy making: Liegmann and van Ackeren (2012) summarize educational reforms following results from PIRLS. Tobin and colleagues (2015) list data examples from Russia that raised policy makers' awareness about the importance of social and school contexts for students' learning. Australia and New Zealand identified a need for teachers' professional development, and Japan highlighted the importance of students' interest and attitudes about learning. Ireland published a report summarizing the results of PIRLS and TIMSS 2001 (Eivers & Clerkin, 2013), including a chapter on teaching practices (Clerkin, 2013), that highlights the importance of professional development and the need for systematic collaboration as school-level policies. The focus on teaching quality in studies such as TIMSS, which is expressed in students' and teachers' judgement in the questionnaires, seems to show in current policy efforts, including the need to participate in future cycles of the assessment to evaluate change (Jones & Bunting, 2013).

ILSAs serve different purposes, and there is a need to report to various stakeholders in education. This paper set out to show that context questionnaire indicators are an important part of ILSAs and that their impact has increased in the past decade. The combination of different theoretical approaches in ILSAs into one overarching framework on a conceptual level is a matter of debate. Moreover, specific indicators and their translation into the measurement level should be addressed, especially when looking at changes over time.

When it comes to the analysis of questionnaire data, traditional scaling models are now being replaced by item response theory (IRT) models, with TIMSS using the Partial Credit Model (Martin et al., 2016) and PISA using the Generalized Partial Credit Model (Buchholz & Jude, 2017) for scaling both cognitive and questionnaire data. However, secondary analysis has already gone beyond these rather unidimensional approaches. Questionnaire indicators can be used to describe different patterns of learning environments across countries (Kobarg et al., 2011; Vieluf et al., 2012). Learning settings can thus be compared across countries by differences in profiles resulting from questionnaire indicators. The use of profiles rather than single-scale comparisons has great potential for reporting and should be considered at the stage of framework development to inform indicator selection for future cycles of ILSA.

For quite some time now, the possibilities and the limitations of drawing causal inferences from cross-sectional ILSAs have been discussed, along with statistical models aiming at estimates of change (see, for example, the special issue on *Large-Scale Assessment in Education*, edited by Rutkowski & Delandshere, 2016). The potential of cross-sectional studies can be significantly enhanced by a strategic implementation of a longitudinal component that needs to include carefully selected context indicators to help explain changes over time (OECD, 2010). More and more countries have taken up this possibility, starting with Canada, which followed the PISA 2000 population (OECD, 2012), and Hong Kong, which has currently implemented a longitudinal component after PISA 2012 (Chinese University of Hong Kong, 2018), to name a couple.

In Germany, several additions to ILSAs have implemented a longitudinal component. For PISA 2003 and PISA 2012, students participated in a repeated test and questionnaire assessment design 1 year after the initial assessment (Reiss et al., 2017). This allowed for a longitudinal analysis of change over one school year, that is, between 2003 and 2004 and between 2012 and 2013, taking into account context factors on the individual and the school level. Differences in cognitive activation and classroom instruction showed an impact on competence development that also differed between boys and girls. Earlier, Germany implemented a school panel (i.e., a repeated assessment on the school level that sampled PISA 2009 students in schools that participated in 2000) and analyzed medium-term changes at the school level. Results showed the impact of internal school evaluation on the respective students' cohort outcomes over time (Bischof et al., 2013). Just recently, an individual panel study was implemented for PIAAC in Germany that takes into account repeated measurement of context factors such as life satisfaction and health (Rammstedt et al., 2017).

To conclude, the question remains about how to cover a broad and increasing range of context indicators without adding to individual assessment time. Besides investing in further research on how to optimize a booklet design for questionnaires (without endangering quality standards in any other aspect of the study), the idea of international options seems promising. By offering short additional questionnaires that cover specific content in more depth, countries can choose which policy areas would be of interest to them, and international comparison would remain possible provided that many countries choose to implement the same option. The concept of policy areas introduced with the modular approach of the PISA 2015 and PISA 2018 frameworks (Jude, 2016; Klieme & Kuger, 2016) could serve as the foundation for the development of indicators in specific areas that would add to the standard indicators assessed by a common core student questionnaire.

ILSA designs can vary regarding their flexibility, and limitations still arise from technical issues, such as the limited testing time allocated to context questionnaires. To our knowledge, there is no empirical argument as to why only 30 to 35 minutes are usually assigned to the questionnaires even though they need to deliver more indicators than the cognitive test that may take up to 2 hours, or why questionnaires are assigned after the cognitive tests (i.e., beyond the burden of an overall testing time of 2.5 hours). We would advocate for a more prominent role of context questionnaires in ILSAs regarding study design, as well as the impact they have on reporting. While the media and policy makers alike are most affected by league tables and ranking of cognitive skills, indicators such as equity, students' well-being, and teacher motivation could arguably be considered as more important because they are prerequisites for learning. Policy makers should definitely take an interest in the questionnaires, which capture the "how and why" students learn.

However, caution is still advised when interpreting the resulting data. Cross-sectional ILSAs do not allow for following the development of individual students or schools over time, and statistical associations need to be treated carefully when they are meant to drive policy decisions. No mechanisms exist for drawing causal inferences from data delivered by ILSAs. Still, ILSAs are highly important for describing trends and developments on a country level and for tracking indicators and their relationship over time.

The more indicators that are available to report on the context of learning, the more knowledge exists that can serve as a basis for policy decisions. Larger trends will only become visible over time if they can rely on a sound indicator system. It is therefore necessary to identify policy-relevant indicators that should find their way into overarching frameworks and be used for continuous measurement. A sound comprehensive documentation of indicators across studies and across years, including searchable meta-data that are linked to the respective datasets, need to be made available for further research. Overall, a broader discussion is needed to shed light on how to identify constructs that are relevant for continually monitoring education and policy making, both internationally and nationally.

REFERENCES

Adams, R. J., Lietz, P., & Berezner, A. (2013). On the use of rotated context questionnaires in conjunction with multilevel item response models. *Large-scale Assessments in Education, 1*(1), 1-5. https://doi.org/10.1186/2196-0739-1-5.

Agudo-Peregrina, Á. F., Iglesias-Pradas, S. Conde-González, M. A., & Hernández-García, A. (2014). Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning. *Computers in Human Behavior, 31*, 542-550.

Allen, J., & van der Velden, R. (2014). Contextual indicators in adult literacy studies: The case of PIAAC. In L. Rutkowski, M. von Davier, M., & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 345-360). Boca Raton, FL: CRC Press.

Almlund, M., Duckworth, A. L., Heckman, J. J., & Kautz, T. D. (2011). *Personality psychology and economics*. Bonn, Germany: IZA. http://nbn-resolving.de/urn:nbn:de:101:1-201104113733.

Almond, R., Deane, P., Quinlan, T., Wagner, M., & Sydorenko, T. (2012). A preliminary analysis of keystroke log data from a timed writing task. Research Report ETS RR-12-23. *ETS Research Report Series*. https://www.ets.org/Media/Research/pdf/RR-12-23.pdf.

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

Bartram, D. (2007). Increasing validity with forced-choice criterion measurement formats. *International Journal of Selection and Assessment, 15*, 263-272. doi:10.1111/j.1468-2389.2007.00386.x.

Bertling, J. P., Borgonovi, F., & Almonte, D. E. (2016). Psychosocial skills in large-scale assessments: Trends, challenges, and policy implications. In A. Lipnevich, F. Preckel, & R. Roberts (Eds.), *Psychosocial skills and school systems in the 21st century. The Springer series on human exceptionality* (pp. 347-372). Cham, Switzerland: Springer International Publishing. https://doi.org/10.1007/978-3-319-28606-8_14.

Bischof, L. M., Hochweber, J., Hartig, J., & Klieme, E. (2013). Schulentwicklung im Verlauf eines Jahrzehnts. Erste Ergebnisse des PISA-Schulpanels. In N. Jude & E. Klieme (Eds.), *PISA 2009 - Impulse für die Schul - und Unterrichtsforschung,* 172-199. *Zeitschrift für Pädagogik,* (Beiheft) 59. Weinheim u.a.: Beltz.

Blom, A.G., Jäckle, A., & Lynn, P. (2010). The use of contact data in understanding cross-national differences in unit nonresponse. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennel, & T. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 335-354). Sussex, UK: Wiley.

Borgonovi, F., & Pál, J. (2016). A framework for the analysis of student well-being in the PISA 2015 study: Being 15 in 2015. *OECD Education Working Papers*. doi:10.1787/199.

Breakspear, S. (2014). *How does PISA shape education policy making? Why how we measure learning determines what counts in education.* Centre for Strategic Education Seminar Series Paper No. 240, November 2014.

Brese, F., & Mirazchiyski, P. (2013). Measuring students' family background in large-scale international education studies. *IERI Monograph Series. Issues and Methodologies in Large-Scale Assessments*. Hamburg, Germancy: ETS/IEA.

Brunello, G., & Schlotter, M. (2011). *Non-cognitive skills and personality traits: Labour market relevance and their development in education and training systems* (IZA Discussion Paper No. 5743). https://ssrn.com/abstract=1858066.

Buchholz, J., & Jude, N. (2017). Scaling procedures and construct validation of context questionnaire data. *PISA 2015 Technical Report* (pp. 283-315). Paris, France: OECD Publishing.

Buchmann, C. (2002). Measuring family background in international studies of education: Conceptual issues and methodological challenges. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 150-197). Washington, DC: National Academy Press.

Buerger, S., Kroehne, U., & Goldhammer, F. (2016). The transition to computer-based testing in large-scale assessments: Investigating (partial) measurement invariance between modes. *Psychological Test and Assessment Modeling, 58*, 597-616.

Caro, D. H., Lenkeit, J., & Kyriakides, L. (2016). Teaching strategies and differential effectiveness across learning contexts: Evidence from PISA 2012. *Studies in Educational Evaluation, 49*, 30-41. https://doi.org/10.1016/j.stueduc.2016.03.005.

Chinese University of Hong Kong. (2018). *Longitudinal Study of Adolescents in Hong Kong (HKLSA)*. http://www.fed.cuhk.edu.hk/~hkcisa/hklsa.html.

Clerkin, A. (2013). Teachers and teaching practices. In E. Eivers & A. Clerkin ↱↱(Eds.), *National schools, international contexts: Beyond the PIRLS and TIMSS ↱test results* (pp. 77-104). Dublin, Ireland: Educational Research Centre. ↱

Collins, D. (2003). Pretesting survey instruments: An overview of cognitive methods. *Quality of Life Research, 12*(3), 229-238. https://doi.org/10.1023/A:1023254226592.

Commission of the European Commission. (2000). *A memorandum on lifelong learning* (SEC No. 1832). Brussels, Belgium.

Cope, B., & Kalantzis, M. (2016). Big data comes to school: Implications for learning, assessment, and research. *AERA Open, 2*(2), 1-19.

Creemers, B. P. M., & Kyriakides, L. (2010). School factors explaining achievement on cognitive and affective outcomes: Establishing a dynamic model of educational effectiveness. *Scandinavian Journal of Educational Research, 54*(3), 263-294. https://doi.org/10.1080/00313831003764529.

Csapó, B., & Funke, J. (Eds.). (2017). *The nature of problem solving: Using research to inspire 21st century learning*. Paris, France: OECD Publishing.

De Neufville, J. I. (1978). Validating policy indicators. *Policy Sciences, 10*(2-3), 171-188. https://doi.org/10.1007/BF00136034.

Décieux, J., Mergener, A., Neufang, K., & Sischka, P. (2015). Implementation of the forced answering option within online surveys. Do higher item response rates come at the expense of participation and answer quality? *Psihologija, 48*(4), 311-326. doi:10.2298/PSI1504311D.

Dept, S., Ferrari, A., & Halleux, B. (2017). Translation and cultural appropriateness of survey material in large-scale assessments. In P. Lietz, J. C. Cresswell, K. F. Rust, & R. J. Adams (Eds.), *Implementation of Large-Scale Education Assessments* (pp. 168-192). Sussex, UK: Wiley

Dept, S., Ferrari, A., & Wyrinen, L. (2010). Development in translation verification procedures in three multilingual assessments: A plea for an integrated translation and adaptation monitoring tool. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennel, & T. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 157-176). Sussex, UK: Wiley.

Ebbs, D., & Djekić, M. (2016). Translation and translation verification for TIMSS Advanced 2015. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS Advanced 2015* (pp. 7.1-7.10). http://timssandpirls.bc.edu/publications/timss/2015-a-methods/chapter-7.html.

Efrati, V., Limongelli, C., & Sciarrone, F. (2014). A data mining approach to the analysis of students' learning styles in an e-learning community: A case study. In C. Stephanidis & M. Antona (Eds.), *Universal access in human-computer interaction.* UAHCI 2014. Lecture Notes in Computer Science (Vol. 8514). Cham, Switzerland: Springer.

Eigenhuis, A., Kamphuis, J. H., & Noordhof, A. (2017). Personality in general and clinical samples: Measurement invariance of the multidimensional personality questionnaire. *Psychological Assessment, 29*(9), 1111-1119.

Eivers, E., & Clerkin A. (Eds.). (2013). *National schools, international contexts: Beyond the PIRLS and TIMSS test results*. Dublin, Ireland: Educational Research Centre.

Erikson, R., Goldthorpe J. H., & Portocarero, L. (1979). Intergenerational class mobility in three western European societies. *British Journal of Sociology, 30*, 415-441.

European Commission. (2017). Call for tenders. N° EAC/22/2017. *Study on engagement and achievement of 15-year-olds in PISA 2015 across E.U. member states*. Ref. Ares (2017)3638694-19/07/2017.

Fadel, C., Bialik, M., & Trilling, B. (2015). *Four-dimensional education*. Boston, MA: Center for Curriculum Redesign.

Gorges, J., Koch, T., Maehler, D. B., & Offerhaus, J. (2017). Same but different? Measurement invariance of the PIAAC motivation to learn scale across key socio-demographic groups. *Large-scale Assessment Education, 5*(13). https://doi.org/10.1186/s40536-017-0047-5.

Granda, P., Wolf, C., & Hadorn, R. (2010). Harmonizing survey data. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennel, & T. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 315-332). Sussex, UK: Wiley.

Grisay, A., de Jong, J. H. A. L., Gebhardt, E., Berezner, A., & Halleux-Monseur, B. (2007). Translation equivalence across PISA countries. *Journal of Applied Measurement, 8*(3), 249-266.

Gutman, L. M., & Schoon, I. (2013). *The impact of non-cognitive skills on outcomes for young people. Literature review*. London, UK: Institute of Education.

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M., & Steiner, S. (2017). Three methods for occupation coding based on statistical learning. *Journal of Official Statistics, 33*(1), 101-122. https://doi.org/10.1515/JOS-2017-0006.

Hanushek, E. A., Peterson, P. E., & Woessmann, L. (2012). Achievement growth: International and U.S. state trends in student performance. *PEPG Report, 12*(3). Cambridge, MA: Harvard University.

Harkness, J. A., Edwards, B., Hansen, S. E., Miller, D. R., & Villar, A. (2010a). Designing questionnaires for multipopulation research. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennel, & T. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 33-57). Sussex, UK: Wiley.

Harkness, J. A., Villar, A., & Edwards, B. (2010b). Translation, adaptation, and design. In J. A. Harkness, M. Braun, B. Edwards, T. P. Johnson, L. Lyberg, P. Ph. Mohler, B.-E. Pennel, & T. Smith (Eds.), *Survey methods in multinational, multiregional, and multicultural contexts* (pp. 117-140). Sussex, UK: Wiley.

Harlow, L. L., & Oswald, F. L. (2016). Big data in psychology: Introduction to the special issue. *Psychological Methods, 21*(4), 447-457.

Hartley, S. L., & MacLean, W. E. (2006). A review of the reliability and validity of Likert-type scales for people with intellectual disability. *Journal of Intellectual Disability Research, 50*(Pt 11), 813-827. https://doi.org/10.1111/j.1365-2788.2006.00844.x.

Hattie, J., & Anderman, E. M. (Eds.). (2013). *International guide to student achievement*. New York: Routledge.

He, J., Buchholz, J., & Klieme, E. (2017). Effects of anchoring vignettes on comparability and predictive validity of student self-reports in 64 cultures. *Journal of Cross-Cultural Psychology, 48*(3), 319-334. https://doi.org/10.1177/0022022116687395.

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with N-grams: Insights from a computer-based large-scale assessment. In R. Yigal, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 749-776). Hershey, PA: IGI Global.

Heckman, J. J., & Kautz, T. (2013). *Fostering and measuring skills: Interventions that improve character and cognition* (IZA Discussion Paper No. 7750). Bonn, Germany: Institute for the Study of Labor.

Heckman, J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics, 24*(3), 411-482.

Hershkovitz, A., & Nachmias, R. (2009). Learning about online learning processes and students' motivation through Web usage mining. *Interdisciplinary Journal of E-Learning and Learning Objects,* 5.

Heyneman, S. P., & Lee, B. (2014). The impact of international studies of academic achievement on policy and research. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 37-75). Boca Raton, FL: CRC Press.

Heyneman, S. P., & Lee, B. (2015). *International large-scale assessments.* In H. F. Ladd & M. E. Goertz (Eds.), *Handbook of research in education, finance and policy* (pp. 105-123). Routledge.

Hopfenbeck, T. N., & Maul, A. (2011). Examining evidence for the validity of PISA learning strategy scales based on student response processes. *International Journal of Testing, 11*(2), 95-121.

Hopkins, D., & King, G. (2010). Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability. *Public Opinion Quarterly, 74*(2), 201-222.

Huebener, M., Kuger, S., & Marcus, J. (2017). Increased instruction hours and the widening gap in student performance. *Labour Economics, 47*, 15-34. doi:10.1016/j.labeco.2017.04.007.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. *Journal of Cross-Cultural Psychology, 20*(3), 296-309.

Humphrey, N., Kalambouka, A., Wigelsworth, M., Lendrum, A., Deighton, J., & Wolpert, M. (2011). Measures of social and emotional skills for children and young people. *Educational and Psychological Measurement, 71*(4), 617-637. doi:10.1177/0013164410382896.

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

International Association for the Evaluation of Educational Achievement (IEA). (2000). *Framework and specifications for PIRLS assessment 2001*. Chestnut Hill, MA: Boston College.

International Labour Office (ILO). (2012). *International standard classification of occupations. ISCO-08*. Geneva, Switzerland: ILO.

Jerrim, J., & Micklewright, J. (2014). Socio-economic gradients in children's cognitive skills: Are cross-country comparisons robust to who reports family background? *European Sociological Review, 30*(6), 766-781.

John, O. P., & DeFruyt, F. (2015). *Framework for the longitudinal study of social and emotional skills in cities*. Paris, France: OECD Publishing.

Jones, A., & Bunting, C. (2013). International, national, and classroom assessment: Potent factors in shaping what counts in school science. In D. Corrigan, R. Gunstone, & A. Jones (Eds.), *Valuing assessment in science education: Pedagogy, curriculum, policy* (pp. 33-53). Dordrecht, Netherlands: Springer.

Jude, N. (2016). The assessment of learning contexts in PISA. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning: An international perspective* (pp. 39-51). Dordrecht, Netherlands: Springer.

Kaplan, D. (2016). Causal inference with large-scale assessments in education from a Bayesian perspective: A review and synthesis. *Large-scale Assessment Education, 4*(7), 1-19. https://doi.org/10.1186/s40536-016-0022-6.

Kaplan, D., & Elliott, P. R. (1997). A model-based approach to validating education indicators using multilevel structural equation modeling. *Journal of Educational and Behavioral Statistics, 22*(3), 323-347.

Kaplan, D., & Kreisman, M. B. (2000). On the validation of indicators of mathematics education using TIMSS: An application of multilevel covariance structure modeling. *International Journal of Educational Policy, Research, and Practice, 1*, 217-242.

Kaplan, D., & McCarty, A. T. (2013). Data fusion with international large-scale assessments: A case study using the OECD PISA and TALIS surveys. *Large-scale Assessments in Education, 1*(6). https://doi.org/10.3102/1076998615622221.

Kaplan, D., & Su, D. (2016). On matrix sampling and imputation of context questionnaires with implications for the generation of plausible values in large-scale assessments. *Journal of Educational and Behavioral Statistics, 41*(1), 57-80. https://doi.org/10.3102/1076998615622221.

King, G., & Wand, J. (2007). Comparing incomparable survey responses: New tools for anchoring vignettes. *Political Analysis, 15*, 46-66.

Klieme, E. (2016). *TIMSS 2015 and PISA 2015. How are they related on the country level?* DIPF working paper. https://www.dipf.de/de/forschung/publikationen/pdf-publikationen/Klieme_TIMSS2015and PISA2015.pdf.

Klieme, E., & Kuger, S. (2016). PISA 2015 context questionnaires framework: Monitoring opportunities and outcomes, policies and practices modelling patterns and relations, impacts and trends in education. In Organisation for Economic Co-operation and Development (OECD) (Ed.), *PISA. PISA 2015 assessment and analytical framework* (pp. 101-127). OECD Publishing. https://doi.org/10.1787/9789264255425-7-en.

Kobarg, M., Prenzel, M., Seidel, T., Walker, M., McCrae, B., Cresswell, J., & Wittwer, J. (2011). *An international comparison of science teaching and learning. Further results from PISA 2006*. Münster, Germany: Waxmann.

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

Kunz, T., & Fuchs, M. (2018). Dynamic instructions in check-all-that-apply questions. *Social Science Computer Review*. doi:10.1177/0894439317748890.

Kyllonen, P. C., & Bertling, J. P. (2013). Innovative questionnaire assessment methods to increase cross-country comparability. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 277-285). Boca Raton, FL: CRC Press.

Laschke, C., & Blömeke, S. (2016). Measurement of job motivation in TEDS-M: Testing for invariance across countries and cultures. *Large-scale Assessments in Education, 4*(16). https://doi.org/10.1186/s40536-016-0031-5.

Lee, J. W., Jones, P. S., Mineyama, Y., & Zhang, X. E. (2002). Cultural differences in responses to a Likert Scale. *Research in Nursing and Health, 25*, 295-306. doi:10.1002/nur.10041.

Leiulfsrud, H., Bison, I., & H. Jensberg (2005). *Social class in Europe*. *European Social Survey* 2002/3. Trondheim, Norway: NTNU Social Research, Ltd.

Lenkeit, J., & Caro, D. H. (2014). Performance status and change—Measuring education system effectiveness with data from PISA 2000–2009. *Educational Research and Evaluation, 20*(2), 146-174. http://dx.doi.org/10.1080/13803611.2014.891462.

Liegmann, A. B., & van Ackeren, I. (2012). The impact of PIRLS in 12 countries: A comparative summary. In K. Schwippert & J. Lenkeit (Eds.), *Progress in reading literacy in national and international context. The impact of PIRLS 2006 in 12 countries* (pp. 228-252). Münster, Germany: Waxmann.

Lietz, P. (2017). Design, development and implementation of contextual questionnaires. In P. Lietz, J. C. Cresswell, K. F. Rust, & R. J. Adams (Eds.), *Large-scale assessments, in implementation of large-scale education assessments* (pp. 92-136). Chichester, UK: Wiley & Sons, Ltd. doi:10.1002/9781118762462.ch4.

Lindqvist, E., & Vestman, R. (2011). The labor market returns to cognitive and noncognitive ability: Evidence from the Swedish enlistment. *American Economic Journal: Applied Economics, 3*(1), 101-128. https://doi.org/10.1257/app.3.1.101.

Marksteiner, T., & Kuger, S. (2016). Sense of belonging to school in 15-year-old students—The role of parental education and students' attitudes toward school. *European Journal of Psychological Assessment, 32*(1), 68-74.

Marksteiner, T., Kuger, S., & Klieme, E. (in review). The potential of anchoring vignettes to increase intercultural comparability of non-cognitive factors. *Assessment in Education: Principles, Policy & Practice.*

Martens, K., Niemann, D., & Teltemann, J. (2016). Effects of international assessments in education—A multidisciplinary review. *European Educational Research Journal, 15*(5), 516-522.

Martin, M. O., Mullis, I. V. S., Gonzalez, E. J., Smith, T. A., & Kelly, D. L. (1999). *School contexts for learning and instruction: IEA's third International Mathematics and Science Study*. Chestnut Hill, MA: Boston College.

Martin, M. O., Mullis, I. V. S., Hooper, M., Yin, L., Foy, P., & Palazzo, L. (2016). Creating and interpreting the TIMSS 2015 context questionnaire scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 15.1-15.312). Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center. http://timss.bc.edu/publications/timss/2015-methods/chapter-15.html.

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 Assessment in Education, 5*(3). https://doi.org/10.1186/s40536-017-0038-6.

Mohadjer, L., Krenzke, T., & Van de Kerckhove, W. (2013). Indicators of the quality of the sample data. In OECD (Ed.), *Technical report of the survey of adult skills (PIAAC)* (pp. 1-30). Paris, France: OECD Publishing.

Mullis, I. V. S., & Martin, M. O. (Eds.). (2013). *TIMSS 2015 assessment frameworks.* http://timssandpirls.bc.edu/timss2015/frameworks.html.

Mullis, I. V. S., Martin, M. O., Ruddock, G. J., O'Sullivan, C. Y., Arora, A., & Erberber, E. (2007). *TIMSS 2007 assessment frameworks.* Chestnut Hill, MA: Boston College.

Mullis, I. V. S., Martin, M. O., Smith, T. A., Garden, R. A., Gregory, K. D., Gonzalez, E. J., Chrostowski, S. J., & O'Connor, K. M. (2001). *TIMSS assessment frameworks and specifications 2003.* Chestnut Hill, MA: Boston College.

Nagengast, B., & Marsh, H. W. (2014). Motivation and engagement in science around the globe: Testing measurement invariance with multigroup structural equation models across 57 countries using PISA 2006. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 317-344). Boca Raton, FL: CRC Press.

National Academy of Education (NAEd). (2017). *Big data in education: Balancing the benefits of educational research and student privacy: Workshop summary*. Washington, DC: National Academy of Education.

National Research Council (NRC). (1999). *Grading the nation's report card: Evaluating NAEP and transforming the assessment of educational progress.* Washington, DC: National Academy Press.

Naumann, J. (2015). A model of online reading engagement: Linking engagement, navigation, and performance in digital reading. *Computers in Human Behavior, 53*, 263-277.

O'Connell, M. S., Hartman, N. S., McDaniel, M. A., Grubb, W. L., & Lawrence, A. (2007). Incremental validity of situational judgment tests for task and contextual job performance. *International Journal of Selection and Assessment, 15*, 19-29. doi:10.1111/j.1468-2389.2007.00364.x.

Organisation for Economic Co-operation and Development (OECD). (1999). *Measuring student knowledge and skills: A new framework for assessment*. Paris, France: OECD Publishing.

OECD. (2010). *Pathways to success: How knowledge and skills at age 15 shape future lives in Canada*. Paris, France: OECD Publishing. http://dx.doi.org/10.1787/9789264081925-2-en.

OECD. (2012). *Learning beyond fifteen: Ten years after PISA*, Paris, France: OECD Publishing. http://dx.doi.org/10.1787/9789264172104-en.

OECD. (2014). *PISA 2012 technical report. Context questionnaire development*. Paris, France: OECD Publishing.

OECD. (2015). *PISA 2015 technical report.* Paris, France: OECD Publishing.

OECD. (2016a). *Insights from the TALIS-PISA link data: Teaching strategies for instructional quality.* Paris, France: OECD Publishing.

OECD. (2016b). *PISA 2015 technical report*. Paris, France: OECD Publishing.

OECD. (2017a). *PISA 2015 results (Volume V): Collaborative problem solving*. Paris, France: OECD Publishing.

OECD. (2017b). *PISA 2015 results (Volume III): Students' well-being*. Paris, France: OECD Publishing.

OECD. (2017c). *Education at a glance 2017. OECD indicators.* Paris, France: OECD Publishing.

Ortmanns, V., & Schneider, S. L. (2015). Harmonization still failing? Inconsistency of education variables in cross-national public opinion surveys. *International Journal of Public Opinion Research*. doi:10.1093/ijpor/edv025.

Papamitsiou, Z., & Economides, A. A. (2017). Exhibiting achievement behavior during computer-based testing: What temporal trace data and personality traits tell us? *Computers in Human Behavior, 75*, 423-438. https://doi.org/10.1016/j.chb.2017.05.036.

Patterson, F., Ashworth, V., Zibarras, L., Coan, P., Kerrin, M., & O'Neill, P. (2012). Evaluations of situational judgement tests to assess non-academic attributes in selection. *Medical Education*, *46*(9), 850-868. https://doi.org/10.1111/j.1365-2923.2012.04336.x.

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology, 84*(4), 890-904. http://dx.doi.org/10.1037/0022-3514.84.4.890.

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

Purves, A. C. (1987). The evolution of the IEA: A memoir. *Comparative Education Review, 31*(1), 10-28. https://doi.org/10.1086/446653.

Rammstedt, B., Martin, S., Zabal1, A., Carstensen, C., & Schupp, J. (2017). The PIAAC longitudinal study in Germany: Rationale and design. *Large-scale Assessment Education, 5*(4). https://doi.org/10.1186/s40536-017-0040-z.

Reiss, K., Klieme, E., Köller, O., & Stanat, P. (2017). PISA Plus 2012–2013: Kompetenzentwicklung im Verlauf eines Schuljahres, *20*(2), Supplement.

Richardson, M., Abraham, C., & Bond, R. (2012). Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological Bulletin, 138*(2), 353-387.

Romero, C., & Ventura, S. (2010). Educational data mining: A review of the state of the art. IEEE transactions on systems, man, and cybernetics—Part C: Applications and Reviews. Vol. 40, No. 6.

Rutkowski, D., & Delandshere, G. (2016). Causal inferences with large-scale assessment data: Using a validity framework. *Large-scale Assessment Education, 4*(6). https://doi.org/10.1186/s40536-016-0019-1.

Rutkowski, L., & Rutkowski, D. (2010). Getting it "better": The importance of improving background questionnaires in international large-scale assessment. *Journal of Curriculum Studies, 42*(3), 411-430. http://dx.doi.org/10.1080/00220272.2010.487546.

Rutkowski, D., & Rutkowski, L. (2013). Measuring socioeconomic background in PISA: One size might not fit all. *Research in Comparative and International Education, 8*(3), 259-278.

Rutkowski, L., & Rutkowski, D. (2017). Improving the comparability and local usefulness of international assessments: A look back and a way forward. *Scandinavian Journal of Educational Research, 2*(3), 1-14. https://doi.org/10.1080/00313831.2016.1261044.

Rutkowski, L., & Svetina, D. (2017). Measurement invariance in international surveys: Categorical indicators and fit measure performance. *Applied Measurement in Education, 30*(1), 39-51.

Rychen, D. S., & Salganik, L. H. (2003). *Key competencies for a successful life and a well-functioning society*. Cambridge, MA: Hogrefe & Huber.

Scheerens, J. (2015). Theories on educational effectiveness and ineffectiveness. *School Effectiveness and School Improvement, 26*(1), 10-31. https://doi.org/10.1080/09243453.2013.858754.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness.* Oxford, UK: Pergamon.

Schmidt, W. H., Burroughs, N. A., Zoido, P., & Houang, R. T. (2015). The role of schooling in perpetuating educational inequality. *Educational Researcher, 44*(7), 371-386. https://doi.org/10.3102/0013189X15603982.

Schulz, W., Ainley, J., Fraillon, J., Losito, B., & Agrusti, G. (2016). *IEA International Civic and Citizenship Education Study 2016 assessment framework.* Amsterdam, Netherlands: IEA.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417-453.

Synergies for Europe's Research Infrastructures in the Social Sciences (SERISS). (2017). *A coding module for socio-economic survey questions*. https://seriss.eu/about-seriss/work-packages/wp8-a-coding-module-for-socio-economic-survey-questions.

Tawil, S. (2013). *Education for "global citizenship": A framework for discussion*. ERF Working Papers Series, no. 7. Paris, France: UNESCO.

Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research.* London, UK: Routledge Falmer.

Teltemann, J., & Klieme, E. (2016). The impact of international testing projects on policy and practice. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of social and human conditions in assessment* (pp. 1-39). New York: Routledge.

Tijdens, K. (2015). *The design of a tool for the measurement of occupations in Web surveys using a global index of occupations.* Working paper, Leuven, InGRID project, M21.4. http://www. Inclusivegrowth.be/project-output.

Tobin, M., Lietz, P., Nugroho, D., Vivekanandan, R., & Nyamkhuu, T. (2015). *Using large-scale assessments of students' learning to inform education policy: Insights from the Asia-Pacific region*. Melbourne, Australia: ACER and Bangkok, Thailand: UNESCO.

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times.* Sussex, UK: Jossey-Bass.

van de Vijver, F. J. R, & He, J. (2016). Bias assessment and prevention in non-cognitive outcome measures in PISA questionnaires. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning: An international perspective* (pp. 229-253). Cham, Switzerland: Springer.

van de Vijver, F. J. R, Jude, N., & Kuger, S. (in review). *Challenges in international large-scale educational surveys*. In L. Sage (Ed.), *Handbook on comparative international studies*.

van de Vijver, F. J. R., & Leung, K. (1997). Methods and data analysis of comparative research (second ed.). In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology, 1* (pp. 257-300). Boston, MA: Allyn & Bacon.

Vieluf, S., Kaplan, D., Klieme, E., & Bayer, S. (2012). *Teaching practices and pedagogical innovation: Evidence from TALIS*. Paris, France: OECD Publishing. https://doi.org/10.1787/9789264123540-en.

Walton, G. M., & Cohen, G. L. (2011). A brief social-belonging intervention improves academic and health outcomes of minority students. *Science, 331*, 1447-1451.

Watermann, R., Maaz, K., Bayer, S., & Roczen, N. (2016). Social background. In S. Kuger, E. Klieme, N. Jude, & D. Kaplan (Eds.), *Assessing contexts of learning: An international perspective* (pp. 117-145). Cham, Switzerland: Springer.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91*(3), 461-481.

Williamson, B. (2017). *Big data in education. The digital future of learning, policy and practice*. London, UK: SAGE.

Willms, J. D., & Tramonte, L. (2014). *Towards the development of contextual questionnaires for the PISA for development study.* http://www.oecd.org/callsfortenders/Annex%20E%20-%20Contextual%20Questionnaires%20Paper.pdf.

Xiao, Y., Liu, H., & Li, H. (2017). Integration of the forced-choice questionnaire and the Likert Scale: A simulation study. *Frontiers in Psychology, 8*, 806. http://doi.org/10.3389/fpsyg.2017.00806.

Yap, S. C. Y., Donnellan, M. B., Schwartz, S. J., Kim, S. Y., Castillo, L. G., Zamboanga, B. L., Weisskirch, R. S., Lee, R. M., Park, I. J. K., Whitbourne, S. K., & Vazsonyi, A. T. (2014). Investigating the structure and measurement invariance of the Multigroup Ethnic Identity Measure in a multiethnic sample of college students. *Journal of Counseling Psychology, 61*(3), 437-446. https://doi.org/10.1037/a0036253.

Zhu, M., Shu, Z., & von Davier, A. (2016). Using networks to visualize and analyze process data for educational assessment. *Journal of Educational Measurement, 53*(2), 190-211.

Zuzovsky, R. (2013). What works where? The relationship between instructional variables and schools' mean scores in mathematics and science in low-, medium-, and high-achieving countries. *Large-scale Assessments in Education, 1*(1), 2. https://doi.org/10.1186/2196-0739-1-2.