Chapter 5

Questionnaire Development in International Large-Scale Assessment Studies

Wolfram Schulz and Ralph Carstens

Abstract Questionnaires are a crucial part of international large-scale studies of educational achievement, such as the International Association for the Evaluation of Educational Achievement (IEA) studies on mathematics, science, reading, computer and information literacy, and civic and citizenship education. Building on IEA's well-established technical standards, this chapter provides an overview of the purpose of this type of instrument, approaches to its development, and the evolving challenges in this area. While large-scale assessment studies have traditionally employed questionnaires to gather contextual information to explain variation in the respective outcome variables of interest, over time there has been a shift toward also collecting information and reports on students' attitudes, dispositions, or behaviors as outcome measures. More recently, development of alternative item formats and approaches have further increased the reliability, validity, and comparability within and across education systems. Approaches to questionnaire purpose and design can vary and instrument can be targeted to different groups and populations. Contextual information has to be collected across highly diverse educational systems creating issues regarding cross-national validity. The challenges of maximizing measurement invariance across highly diverse national contexts and the opportunities provided by a computer-based delivery may lead to future changes and improvements in the approach to questionnaire development.

Keywords Contextual information • Pre-testing • Questionnaire design • Question types and formats • Response process

W. Schulz (✉)

Australian Council for Educational Research (ACER), Camberwell, Australia

e-mail: wolfram.schulz@acer.org

R. Carstens

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

e-mail: ralph.carstens@iea-hamburg.de

5.1 Introduction

International studies of educational achievement, such as those conducted by IEA, routinely employ questionnaires to gather contextual information that can be used to explain variation in the outcome variables measured by educational achievement tests. These questionnaires are administered to students, teachers, schools, and/or parents, as well as national research coordinators in the participating countries. In some studies, questionnaires (in particular those administered to students) also play a key role in measuring affective-behavioral student learning outcomes in addition to cognitive test results. This important purpose is particularly relevant for IEA's civic and citizenship education studies (see IEA 2020a). There are also international large-scale assessments (ILSAs) that rely exclusively on (school and teacher) questionnaires to provide self-reported data on educational contexts as the main focus of cross-national surveys (such as IEA's Second Information Technology in Education Study [SITES]; see IEA 2020a).

Here we describe and discuss the different approaches regarding the purpose and design of the questionnaires that support ILSAs and the methods used to target these instruments to the most proximal information source, groups, populations and contexts. Differing question formats reflect the cognitive response process, and rigorous development procedures and quality assurance procedures are used to ensure the validity, reliability, and comparability of data. As a consequence, the IEA has published technical standards for its studies (Martin et al. 1999), and there are other salient guidelines and recommendations, especially the Survey Research Center's (2016) cross-cultural survey guidelines or summary volumes of contemporary issues and approaches (e.g., Johnson et al. 2018; Presser et al. 2004). Given the ongoing transition of paper-based to computer-based delivery in ILSAs, there are also implications related to the different modes of administration.

With regard to these different aspects, we discuss implications of recent developments in questionnaire design and reflect on the future role and characteristics of this type of instrument, including possibilities for improving the validity, reliability, and comparability of questionnaire data from international studies.

5.2 Approaches to Questionnaire Design and Framing

Questionnaire instruments in educational research may serve a range of different purposes. In many studies focused on measuring student achievement, the traditional role of questionnaires has been to provide information that may explain variation in the primary survey measures (e.g., mathematics skills or reading abilities). This type of questionnaire focuses on students' characteristics and home background, teachers' reports on learning environments, and/or principals' report on the educational context in which students learn.

Increasingly, other variables have been recognized as important in their own right, such as students' attitudes toward subject areas, their sense of self-efficacy, or self-regulated learning. While these variables may be related to achievement, their relationship is less clear in terms of expectations of possible causality and there is a growing recognition of their importance as outcome variables of interest in their own right. In certain studies, such as the IEA studies on civic and citizenship education (the Civic Education Study [CIVED] and International Civic and Citizenship Education Study [ICCS]; see IEA 2020a), measures derived from questionnaires (regarding civic attitudes and engagement) are as important in their role as learning outcomes as cognitive measures (civic knowledge and understanding). Obviously, questionnaire-only studies may also include important outcome measures, for example teachers' job satisfaction in the case of the Organisation for Economic Co-operation and Development's (OECD's) Teaching and Learning International Survey (TALIS; OECD 2019a), where all primary variables of interest are derived from questionnaire data.

In particular, in earlier IEA studies that were focused on measuring students' cognitive achievement, the development of questionnaires primarily aimed at gathering data about factors that explained variation in test scores. Apart from collecting basic information about student characteristics, such as gender and age, questionnaires of this kind typically aim to gather data about home and school contexts.

The type of information collected for this purpose is selected based on assumptions about what might provide explanatory factors predicting differences in educational achievement, and, consequently hint at aspects of system and school effectiveness. Typically, assessment frameworks for ILSAs not only describe the content of learning outcome domain(s) that should be measured but also include contextual frameworks, which outline the factors regarded as relevant for explaining variation in the learning outcome variables (see Chap. 3 for a more detailed description of the role of assessment frameworks). A secondary purpose of contextual questionnaires of this kind is also that data help to describe contexts in their own right, which is particularly useful in cases where statistical information (e.g., about school resources) is not available. For examples from the most recent cycles of IEA's Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS) (see IEA 2020a), we refer readers to Hooper (2016) and Hooper and Fishbein (2017). For international perspectives with a focus on PISA, we advise readers to consult Kuger et al. (2016).

It is possible to distinguish between factual information about contexts (e.g., the number of books at home as reported by students and/or parents, or the number of computers available at school as indicated by school principals or information and communication technology [ICT] coordinators) and perceptions of learning environment factors (e.g., student reports on student-teacher relations at school, or teacher reports on teacher collaboration at school). Factual information is often reported based on single variables (e.g., the average age of teachers) or variables are combined to create a new derived measure based on simple arithmetic calculations (e.g., the student-teacher ratio at school). Perceptions and attitudinal aspects are often measured through rating-scale items that then are either reported as single variables

or used to derive a scaled index (in the form of raw, factor, or item response theory [IRT] scores).

International studies with a traditionally strong focus on the measurement of student achievement, such as IEA's TIMSS and PIRLS, or the OECD's Programme for International Student Assessment (PISA), primarily collect questionnaire measures that are helpful to explain achievement. Increasingly though, ILSAs also include a number of affective-behavioral outcomes that relate factors such as attitudes toward learning, self-efficacy, or self-regulated learning, which do not always show strong correlations with achievement but help describe learning contexts. In the IEA's International Computer and Information Literacy Study (ICILS; see IEA 2020a), the measurement of students' and teachers' use of, familiarity with, and attitudes toward digital technology have always played an important role in its reporting of study results, and have received attention both in the initial reports and later secondary research (see Fraillon et al. 2014, 2020).

In IEA's studies of civic and citizenship education, measuring variables related to students' attitudes and engagement has traditionally been as important as the measurement of cognitive aspects (Schulz et al. 2010, 2018b; Torney et al. 1975; Torney-Purta et al. 2001). Here, affective-behavioral domains have a similar weight and importance for the reporting of study results and therefore the study places considerable emphasis on the development of questionnaire items that measure both contextual information and important civic-related learning outcomes. In this way, studies related to this learning area have achieved a more balanced representation of: (1) knowledge- and skill-related aspects; (2) attitudes and dispositions; and (3) practice, engagement, and behavioral intentions.

There are also studies that do not gather cognitive data reflecting student achievement where all reported results are derived from questionnaire instruments. Such studies include IEA's SITES 2006, OECD's TALIS 2008, 2013, and 2018 (OECD 2019a), and the OECD TALIS Starting Strong Survey 2018 (OECD 2019b), all of which have conducted surveys of school staff (teachers and/or educators) and management (school principals, ICT coordinators, and center leaders) (for more details see, e.g., Law et al. 2008 for SITES 2006, and Ainley and Carstens 2018 for OECD TALIS 2018). The aim of these studies is to gather and report information about school contexts, pedagogical beliefs, and teaching practices that are reported as criterion variables independently of achievement outcomes.

Each large-scale study and survey in education requires a careful articulation of research interests, aims, and data needs in order to obtain more advanced empirical insights with potential implications for educational policy, and the respective frameworks have important implications for the role assigned to questionnaires. For example, TIMSS 2019 used national curricula as the major organizing concept in considering how educational opportunities are provided to students, and emphasized those factors that influence these provisions and how they are used by students and teachers (Mullis and Martin 2017). In contrast, for example, the framework for ICCS 2016 (Schulz et al. 2016) laid more emphasis on the role of out-of-school contexts

and included perspectives beyond education, such as general perceptions of important issues in society, the economy and environment, and students' engagement in the local and wider community (e.g., through the use of social media).

5.3 Targeting of Questionnaires to Different Groups and a Diversity of Contexts

Consequently, there are a variety of contexts that may be of interest when collecting contextual material for a study. However, some respondents to a survey may not be sufficiently knowledgeable to provide data on all of the contexts of interest or (which is the central idea of this publication) may not be able to provide information and views with the same degree of validity and reliability. For example, while it is relatively straightforward to ask students and teachers about their perceptions of what happens during classroom interactions, school principals could provide broad information about what is expected with regard to policies or teaching practices at the school, but yet be unable to report on their actual implementation and/or the variation in perceptions regarding implementation.

The contexts that are of relevance for a particular study should be defined in the contextual framework together with the scope for gathering relevant data from respondents about aspects of interest. Contexts that may be of relevance in a particular international study can be mapped against a number of particular survey instruments that could be used to collect information about them (Table 5.1).

It is often also possible to ask more indirect questions. For example, school principals may be asked about the social background of enrolled students or their general expectations of teaching practices in classroom contexts. Furthermore, in smaller contexts (for example, in early learning studies), it might be appropriate to ask teachers (or early education center staff) about individual children.

Generally, when measuring contextual aspects, there are differences in how much information respondents may have about particular topics. For example, students (at least from a certain age onwards) and teachers can be expected to provide relatively reliable information about their personal characteristics (such as age or gender). However, judgments by school principals about the social context of the local community or the socioeconomic background of students at their school are likely to be less accurate. Furthermore, information gathered from students or teachers about what happens in the classroom also tends to differ considerably given their differing perspectives, even if the point of reference (the classroom) is identical (Fraser 1982).

The sample design may also have implications for the targeting of contextual questionnaires to contexts. In cases of a grade-based design (as is customary in most IEA studies), it is possible to gather specific data about defined classroom (or course) contexts from sampled students at school. In the case of an age-based design, where students are randomly sampled from a specific age group enrolled at selected schools (as in PISA), it is likely that student data reflect a wider range of experiences

Table 5.1 Mapping contexts to questionnaire types

Contexts	Student questionnaire	Parent questionnaire	Teacher questionnaire	School questionnaire	National contexts survey
Individual student	×	×			
Individual teacher			×		
Student peer context	×				
Student home context	×	×			
Classroom context	×		×		
School context	×	×		×	
Community context		×	×	×	
National context					×

Notes × = group is targeted by this questionnaire

than those from grade-based samples, as students from the same age group may be enrolled in different classrooms or course grade levels, or even study programs.

Having a classroom sample (as, e.g., in TIMSS or PIRLS) that relates to a common entity of pedagogical practice also provides an opportunity to ask the corresponding subject teacher specific questions about their teaching practices (see, e.g., Hooper 2016). However, in countries where a broader subject area is taught as part of different individual subjects (such as chemistry, biology, or physics), it remains challenging to relate information from teachers to student learning outcomes because pedagogical input into the content area may come from more than one subject teacher. This becomes even more difficult in cross-curricular learning areas (such as civic and citizenship education), where related content could be spread out across a larger variety of subjects with an even larger number of teachers who might have contributed in different ways and to differing extents to students' learning outcomes (see, e.g., Schulz and Nikolova 2004). Within countries, such as those with a federal state system or different jurisdictions, several approaches to organizing a learning area's curriculum may coexist (see European Commission/EACEA/Eurydice 2017).

One particular challenge related to questionnaire development is the inclusion of aspects that are relevant in many but not all countries. For example, questions about differences across study programs within schools may be of relevance in particular countries (such as Belgium or the Netherlands) but not in those where all students follow the same program in the grade under study (such as Australia, Finland, or Slovenia). To allow countries to pursue particular research as part of an international study, sometimes questionnaire sections are developed as international options that

are only included in those countries where these are regarded as relevant. To ensure the comparability of other (core) questionnaire data, there are limits to how much additional item material can be added. Furthermore, even if administered after the core questions overly long questionnaire may affect the response rates in countries participating in an option. A review of the extent to which optional material can be appropriately added to a study would ideally be part of an international field trial.

There are also cases where item material is only relevant to a sub-group of the target sample, for example for teachers of a particular subject or subject area. Here, questionnaires may include a filter question for teachers regarding their subject or subject area after which only certain teachers are asked to complete a section with questions relevant to their subject area. An example of such an approach can be found in IEA's ICCS, where all teachers teaching at the target grade are surveyed but only those teaching civic-related content are presented with more specific questions related to this particular learning area (see Agrusti et al. 2018).

Aspects of learning areas or subjects under study may also differ across geographic regions with a common historical, cultural, political, and/or educational context. Therefore, certain aspects may be relevant in one particular region but not in another. ICCS addresses this by including regional questionnaires that are developed in close cooperation with national research centers and their experts in these regions. In ICCS 2009, there were regional instruments for countries in Asia, Europe, and Latin America (Kerr et al. 2011), while ICCS 2016 administered regional questionnaires in Europe and Latin America (Agrusti et al. 2018). These regional instruments included civic-related content that was regarded as of particular importance for the respective geographic region or was related to specific aspects that would not have been appropriate for countries outside that region. For example, the European questionnaires in ICCS 2009 and 2016 measured perceptions related to the European Union or cooperation between European countries (Kerr et al. 2010; Losito et al. 2018) while the Latin American instruments focused on issues related to government, peaceful coexistence, and diversity (Schulz et al. 2011, 2018a).

For the successful development of questionnaires in international studies it is important to clearly define the scope of the survey with regard to the targeting of relevant aspects for measurement and appropriate sources of information. As part of the planning of the survey, instruments and the type of questions should be designed so that all aspects can be covered appropriately. It is important to consider which respondents can provide valid and reliable information about the contexts that are seen as relevant in a study. In cross-national studies, it is also crucial to keep in mind that the appropriateness of sources of contextual information may vary across countries. For example, when studying science education in classes, some countries may teach content in one combined subject or in separate subjects, which may require adaptations to the wording or design of subject-related survey instruments or when collecting data from teachers, students, and schools.

5.4 Typology of Questions, Item Formats and Resulting Indicators

As discussed in Sect. 3, a vast array of conceptual considerations, contexts, aims, and reporting needs drive the overall questionnaire design principles and the targeting to particular populations and contexts. The final questions used in the instruments are the primary interface between the aspirations and priorities of researchers working on the development and implementation of comparative surveys in education, and respondents participating in these studies. It needs to be emphasized that questionnaire material in ILSAs is typically delivered through the means of written, self-administered questionnaires. In the case of non-student populations, this is routinely done without the presence of (and possible assistance from) a survey administrator. While many other formats and methods are easy to imagine, including recordings, interview transcripts, or work products, the vast majority of ILSAs rely on the cost-effectiveness of written survey instruments.

Consequently, the questionnaire itself, and perhaps some framing comments made on informational letters, are the only sources of guidance available to respondents regarding the aims of the research, the types of information requested, and instruction to adequately respond to more complex questions. As a consequence, there are possible conflicts between research interests to collect data on complex characteristics, antecedents, inputs, processes, and outcomes and the need to develop and phrase survey questions that can actually be understood by respondents. This tension needs to be resolved in order to maximize the potential yield of valid and reliable information. In international, cross-cultural research, it is also common to encounter issues related to comparability of instruments. More specifically, there is a need to phrase questions in such a way that these can be appropriately translated into different sociocultural contexts and languages. This requirement adds another layer of complexity to the development of questionnaires in ILSAs.

The IEA's technical standards, which were developed at the end of the 1990s to enhance reliability and validity, state that questionnaires should be "clear, simple, concise and manageable" (Martin et al. 1999, p. 43). While the standards do not reflect important developments in survey methodology that have occurred more recently, this premise has not lost its relevance. The standards also request questionnaire development to be specific in terms of the results to be reported; to consider whether the questions will produce credible information, and to review each newly developed question carefully so that it relates to one idea only (i.e., to avoid double-barreled questions); to eschew open-ended questions in the final questionnaires; to ensure that response categories match the question intent, are mutually exclusive, and elicit responses that apply to all of the respondents; to use directions to skip to a later question if they do not apply to respondents; and to arrange questions within an instrument so that their flow is natural and sensible.

In contemporary ILSAs, including all IEA studies, these principles are key to the design of questionnaires, but there are also many other criteria that apply. For example, the use of a specific terminology could be appropriate for questions directed

at one population, such as teachers, but these may not be correctly understood by members of other populations, such as students or parents. Furthermore, seemingly similarly defined and worded terms (such as “students with special education needs”) may be consistently understood within one education system but the terminology may differ for other education systems. Using examples in survey questions with the intention of clarifying certain terms (such as providing “Maths Olympics” as an example when asking principals or teachers about the frequency of “school activities to promote mathematical learning”) may trigger a particular reference or frame of mind but could also potentially narrow the scope of responses to the particular set of examples.

In general, international study center staff working on questionnaire development, associated expert groups, and national research coordinators will need to carefully consider the cognitive processes involved in responding to surveys in order to match research aspirations with the realities of obtaining valid and reliable data. The response process itself may introduce measurement error or bias. Tourangeau et al. (2000) advised that, when asking respondents about information, researchers need to consider the following aspects: the original encoding/acquisition of an experience; the storage of that experience in long-term memory; comprehension of the survey question’s task; retrieval of information from memory; integration/estimation from information retrieved; and mapping and editing a judgment/estimate to the response format. Bias in questionnaire data can be introduced related to each of these aspects, for example through the misunderstanding of a question, lack of memory when asked about relevant information too far back in time, poor estimation of requested information, or deliberate misreporting.

The type of information sought, such as a home context or a teacher’s perception, will drive most of the wording of the corresponding question(s) for which there is a range of different approaches from simple to complex formats. The depth (or richness) of the desired characteristic or process that should be measured will be another criterion for the development of questions (i.e., whether researchers would like to check only the occurrence of an aspect, its frequency and/or intensity, information about its actual workings, or possible impacts).

Factual questions on the existence or frequency of current or very recent characteristics or events, or low inference sociodemographic questions (such as age or gender) for that matter, paired with simple closed response formats have a high probability of yielding valid and reliable information. These can be considered as low inference, namely easily observable or verifiable; appropriate formats would be multiple choice questions with nominal or ordinal response options (e.g., the type of school, frequency of a particular school process) or semi-open formats that require respondents to provide numbers (such as counts of enrolled students at school).

There are many aspects of relevance to ILSAs that cannot be measured in such a direct way. Studying behaviors, attitudes, intentions, or expectations requires the assumption of underlying constructs that cannot be observed directly; these can be termed high inference measures. Instead of formulating direct questions about these constructs, they tend to be measured by administering sets of items to respondents using response (rating) scales that typically have an ordinal level of measurement.

To gather indicators of underlying constructs ILSAs tend to use matrix-type formats to measure dimensions of interest (such as the observed use of ICT during lessons, respondents' sense of self-efficacy, or respondents' interest in learning a particular subject). Commonly used formats are frequency scales with fuzzy quantifiers (such as never, rarely, sometimes, often, or always) to measure frequencies of observations or behaviors, or rating scales reflecting the extent of agreement or disagreement (such as strongly agree, agree, disagree, and strongly disagree) regarding sets of statements relating to the same construct and its various aspects and dimensions.

In light of the model developed by Tourangeau et al. (2000), questions that require a high degree of cognitive burden at each stage and/or relate to more distant events or occurrences would be expected to have a higher probability of introducing measurement error. Similarly, questions of highly personal or sensitive nature may elicit deliberate over- or underreporting in order to preserve the desired self-image of a respondent.

Data for which there are no obvious coding schemes or categorizations, may be initially measured using questions with an open format at the pilot or field trial stage in order to identify appropriate factual response options that are comprehensive and mutually exclusive. Classification of open-ended questions using responses from smaller samples can identify aspects that are not clear at the outset of the development process but relevant from the perspective of respondents.

These examples by no means provide a complete picture of the issues related to finding appropriate questionnaire item formats and contents. Harkness et al. (2016) have further illustrated the richness of the debate, and the options and choices available for the development of questionnaires. Lietz (2010, 2017) and Jude and Kuger (2018) summarized persisting and emerging debates surrounding the design of questionnaires, links to theory, and the overall aim of generating data for educational change. Each study needs to find an appropriate balance between its research aims and aspirations and corresponding practical limitations by using expert judgement in the process of writing new questions, adapting existing questions from validated sources, or refining existing questions. For example, asking students about their parents' or guardians' occupations has been one of the most debated questions in international studies of education, since staff at national research centers generally have to interpret students' (often limited) responses and code these to international standards; it is difficult to evaluate the cost-effectiveness of this process.

Critical debates surrounding best practice for questionnaire development (see, e.g., Harkness et al. 2016; Lietz 2010) have also focused on the use of even versus odd category numbers for Likert-scales (i.e., whether it is appropriate to include or exclude a neutral midpoint in the response categories), unipolar versus bipolar response scales, the direction of response scales (i.e., should they always run positive to negative, or from most to least frequently used, or always in the direction of the latent construct), the concurrent use of positive and negative statements (often resulting in some effect), and the use of definitions and examples to guide respondents' answers. Here, IEA studies routinely aim to minimize the cognitive burden for respondents and avoid inconsistent question and response option design within and across cycles.

Since 2010, novel and innovative item and question formats in ILSAs have evolved. Developmental work on new formats has been conducted primarily in the context of OECD's PISA (OECD 2014), but to some extent also in IEA's ICCS and OECD's TALIS. Examples of these innovative research activities include experiments using candidate methods for improving the reliability and cross-cultural validity of self-assessment measures, so-called Bayesian Truth serum, topic familiarity, forced choice, anchoring vignettes, and situational judgment tests (SJTs) (see, e.g., Jude and Kuger 2018). For example, the field trial of TALIS 2013 included a specific measure to capture teachers' negative and positive impression management behavior (seen as an indicative of socially desirable responses), the field trial of IEA ICCS included forced choice formats, and the field trial of TALIS 2018 included SJTs.

However, as Jude and Kuger (2018) found, these formats and methods have only been able to demonstrate limited success in increasing the validity, reliability, and comparability of questionnaire measures, and many of these formats have only relatively poor cost-effectiveness. This may be related to ethical concerns (e.g., when using fictitious concepts), the cognitive complexity of some measures (in particular when using anchoring vignettes or forced-choice formats with students), the hypothetical nature of situation (in particular SJTs), increased reading load (SJTs and anchoring vignettes), and recognition that these alternative formats demonstrated limited potential to measure and correct for differential response styles within and across countries. With respect to the examples above, while novel item formats were trialed in TALIS 2013, ICCS 2016, and TALIS 2018, they were not included in the main data collections. While research related to new item formats continues within and outside the field of ILSAs, their usage is currently quite limited and how effectively these formats can augment or replace established questionnaire design formats in the future remains unclear.

An ILSA's success depends on the representation of different types of actors and experts in the drafting process. There appears to be an increasing level of convergence across different studies, which is further facilitated by the sharing of expertise by technical advisory boards (such as the IEA's Technical Executive Group), experienced study center staff, distinguished research experts and experts from international organizations (such as IEA), and international collaboration and exchange; all helps to advance the quality, validity, and reliability of questionnaire measurement.

5.5 Development Procedures, Process and Quality Management

Any development of questionnaires should ideally be grounded in a conceptual framework that describes the aspects that should be measured with such instruments (see Chap. 3 for further discussion). This framework needs to define the range of factors that are of relevance, either in terms of providing explanation for learning

outcomes or in terms of deriving survey outcome variables. Developing a conceptual underpinning for questionnaire development tends to be particularly challenging when a wide range of diverse national educational contexts are involved, as is typically the case in ILSA.

Large-scale assessments that are designed as cyclical to monitor changes over time face a particular challenge when it comes to questionnaire development for each new cycle. There is demand from stakeholders to retain material so the same measures can be repeated and hence study data can inform on how contexts, perceptions, attitudes, or behaviors have changed since the last cycle(s). However, there is also a conflicting demand to include new material that addresses recent developments or improves previously used measures.

Questionnaires should be completed within an appropriate time frame that avoids respondent fatigue or refusal to participate due to overly long instruments. Experiences from previous cycles and international field trial studies are used to determine an appropriate length and this can depend on different factors, for example, whether a questionnaire is administered after a lengthy cognitive assessment of two hours or a relatively short assessment of less than one hour. Typically, questionnaires for students, teachers, and other respondents are expected to take between 30 min to one hour to complete (including additional optional components).

Given these time restrictions on instrument length, it is a challenge reconciling the need to retain “trend” measures with providing sufficient space for newly-developed items that address evolving areas and/or replace questions that may be viewed as outdated. The process for making decisions on the retention of old material or inclusion of new material can become particularly difficult within the context of international studies, where the diversity of national contexts may lead to differing priorities and views on the appropriateness of retention and renewal of item material.

Once a conceptual framework has been elaborated, the procedure for item development should include as many reviews and piloting activities as permitted by the often quite restricted time frames. Ideally, the item development phase should include:

- Expert reviews at various stages (for international studies these should also include national representatives with expertise in the assessed domain);
- Cognitive laboratories and focus group assessments for qualitative feedback;
- Translatability assessments for studies where material needs to be translated from a source version (typically in English) into other languages (as is usually the case in international studies);
- Initial piloting of new item material with smaller samples of respondents (in international studies this should involve as many participating countries as possible); and
- A general field trial (in international studies this should include all participating countries) that provides an empirical basis for item selection.

Piloting activities (either qualitative or quantitative) have a strong focus on the suitability of the new item material, but are often not conducted in all participating countries and tend to be based on smaller convenience samples. The inclusion of

questionnaire material in international field trials, in turn, aims to review the appropriateness of an instrument that broadly resembles the final main survey instrument. While this may not always include all of the retained item material from previous cycles, it often includes both old and new items in conjunction; this enables questionnaire designers to look into associations between constructs and review how well new items developed for already existing scales measure the same underlying constructs as the old material.

As already is often the case with quantitatively oriented piloting activities, for the field trial it may be appropriate to use more than one form in order to trial a broader range of item material, given the constraints in terms of questionnaire length. It is possible to arrange the distribution of item sets so that there is overlap and all possible combinations of scales and items can be analyzed with the resulting data sets (see, e.g., Agrusti et al. 2018). Another advantage provided by administering questionnaire material in different forms is the ability to trial alternative formats. For example, researchers may be interested in finding out whether it is more appropriate to use a rating scale of agreement or a scale with categories reflecting frequencies of occurrence to measure classroom climate (as perceived by students or teachers).

Analyses of field trial data tend to focus on issues such as:

- Appropriateness of instrument length and content for the surveyed age group (e.g., through a review of missing data);
- Scaling properties of questionnaire items designed to measure latent traits (e.g., self-efficacy or attitudes toward learning) using classic item statistics, factor analysis, and item response modeling;
- Comparisons of results from questionnaire items included in (a) previous cycle(s) with those newly developed for the current survey;
- Analyses of associations between contextual indicators and potential outcome variables; and
- Reviews of measurement invariance across national contexts using item response modeling and/or multi-group confirmatory factor analysis.

A variety of factors may affect the comparability of questionnaire response data in cross-national studies, and the formats typically used to gauge respondents' attitudes or perceptions may not always consistently measure respondents' perceptions and beliefs across the different languages and cultures (see, e.g., Byrne and van de Vijver 2010; Desa et al. 2018; Heine et al. 2002; van de Gaer et al. 2012; Van de Vijver et al. 2019). With this in mind, international studies have started to build in reviews of measurement invariance during the development stage (see, e.g., Schulz 2009; Schulz and Fraillon 2011). At the field trial stage in particular, with data collected across all participating countries, this type of analysis may identify a potential lack of measurement invariance at item or scale level prior to inclusion in the main survey.

Another important challenge when developing questionnaires is to avoid questions that cause respondents to give answers that are biased toward giving a positive image of themselves, their work, or their institution. Tendencies to provide socially desirable responses in a survey may also vary across national contexts and can be regarded as a potential source of response bias in international studies (Johnson and

Van de Vijver 2003; Van de Vijver and He 2014). While researchers have proposed scales that were developed to measure a construct of social desirability (see, e.g., Crowne and Marlowe 1960), research has also shown that it is difficult to use them for detection and/or adjustment of this type bias, given that they also measure content that cannot be easily disentangled from style (see, e.g., McCrae and Costa 1983). Therefore, while it is important to acknowledge tendencies to give socially desirable answers to certain types of questions, which should be considered in the process of developing and reviewing a question's validity, there is no agreed way of empirically investigating this as part of piloting activities or a field trial.

In summary, any questionnaire development should ideally undergo multiple quality assurance procedures embedded throughout the process. A clear reference document (framework) that outlines research questions, scope, design, and content relevant for the development of questionnaire material in international studies is of critical importance. A staged process that includes different stages of review by national staff and experts, qualitative and quantitative vetting at the earlier stages (ideally including translatability assessments), and a field trial that allows a comprehensive review of cross-national appropriateness and psychometric quality of the item material provide the best option for thorough evaluation of the item material.

5.6 Questionnaire Delivery

In recent years, the delivery of ILSA questionnaires to different target populations has transitioned from traditional paper-based instruments to the use of computer-based technology and the internet. Throughout the questionnaire development process, the choice of the delivery mode for questionnaires and its design have important implications for pretesting, and adaptation and translation (Survey Research Center 2016). Computer-based delivery also provides additional opportunities to collect and use auxiliary process data from electronically delivered questionnaires, as well as the ability to design instruments that enable matrix-sampling of items.

Paper-and-pencil administration of questionnaires was the only viable option for ILSAs of education during the 20th century, although research into internet-delivered surveys had been conducted during the 1990s in relation to public-opinion, health, or household-based surveys (see, e.g., Couper 2008; Dillman et al. 1998). In ILSAs, questionnaires designed for self-completion were typically administered to students as part of a paper-based test during the same assessment session managed by a common administrator. Other contextual questionnaires delivered to adult populations, such as school principals, teachers, or parents, were truly self-administered on paper at a time and location chosen by the respondents and later returned to a school coordinator or mailed directly to the study center. Questionnaire completion as part of a student assessment session was loosely timed, while the self-administration to an adult population was untimed (i.e., respondents could take as little or as much time as they needed).

With the rapidly growing penetration and availability of computers and internet connectivity in schools and at home in the late 1990s and early 2000s, the conditions for educational surveys also changed. The IEA pioneered and trialed the first web-based data collection as part of SITES 2006 (Law et al. 2008). Here, the mode of data collection matched the study's research framework (i.e., the study investigated how and to what extent mathematics and science teachers were using ICT within and outside the classroom). The study offered online administration of teacher, principal, and ICT-coordinator questionnaires to all participating countries; however, not all of them chose this the primary mode and some opted for a primarily or exclusively paper-based delivery (Carstens and Pelgrum 2009). While some countries made online administration the primary mode of collection, others made it optional, while other countries decided to administer the survey on paper only. Overall, about 72% of all questionnaires were administered online and, in the 17 (out of 22) countries that used online collection, about 88% of all respondents used the online mode. There was very little variation between the different groups of respondents (e.g., teachers and principals) but choice of delivery mode differed considerably by other characteristics, in particular across age groups.

Furthermore, SITES 2006 investigated issues of measurement invariance across modes using a split-sample design at the field trial stage. As expected, based on prior research findings, the study observed no major differences in response behavior, styles, non-response, or completion time. Regardless of the delivery mode, questionnaires were self-administered without the presence of an administrator, which is viewed as a key factor explaining differences in response behavior (Tourangeau et al. 2000).

The SITES study and other studies, such as the first cycle of OECD's TALIS in 2008 (a survey implemented by IEA), and IEA's ICCS 2009, paved the way for further work in the area and yielded important insights for the design and administration of online questionnaires accompanied by an alternative paper-based delivery. For example, studies had to find efficient and effective ways to manage the instrument production processes, including adaptation, translation, and verification, without duplicating work (and hence duplicating the chance of errors) for international and national study centers planning to administer paper and online questionnaires side by side.

This paradigm shift has also raised important questions regarding the instrument layout. When using dual delivery modes, obtaining comparable data across the two modes is essential. However, this does not necessarily require an identical design and presentation in both modes, which would be a rather challenging and, possibly impossible endeavor. For example, ILSAs using a computer-based delivery typically present one question at a time in online mode, whereas paper instruments might include multiple (albeit short) questions on one single page.

Skipping logic in online mode has the potential of reducing the response burden further by taking respondents directly to the next applicable question rather than relying on the respondent to omit irrelevant questions. Additional validation and review options can be included, such as a hyperlinked table of contents or checks for plausible number ranges and formats. Furthermore, in cases where a dual mode

is available within the same country (as is often the case in online questionnaires for school principals, teacher, or parents), respondents have the option of requesting or accessing paper versions of questionnaires instead of completing them online, a technical standard aimed at preventing respondents from being excluded because of technical requirements.

As time progresses, access to the internet and the number of respondents able to complete questionnaires online is expected to grow. Correspondingly, across different cycles of IEA studies (see the respective technical reports for TIMSS, PIRLS, ICCS, and ICILS; IEA 2020b), the uptake of online delivered questionnaires has generally increased. For example, while in ICCS 2009 only five out of 38 participating countries opted for online delivery, in ICCS 2016, 16 out of 24 countries selected this option. Finally, the technical design of the questionnaire delivery systems used in these studies make no assumptions or requirements about a particular device, internet browser make or type, or available auxiliary software (such as JavaScript), allowing unrestricted access to online questionnaires by reducing or eliminating technical hurdles. However, issues of confidentiality, security, and integrity of online collected data have started to play an increasingly important role in recent years, in response to public concerns and tightened legal standards.

The shift of the primary collection mode for questionnaires, and later assessments, from paper-based to computer-based delivery has introduced two important opportunities of high relevance. First, computer-based/online delivery enables the collection of process and para-data that can facilitate important insights into the quality of question materials and resulting responses through an analysis of response behavior. Second, electronic delivery potentially enables a more targeted delivery of questionnaires, which could lead to improvements to the so far relatively simple rotational approaches used by some ILSAs, for example, at the field trial stage in ICCS, ICILS, and OECD's TALIS (Agrusti et al. 2018; Ainley and Schulz 2020; Carstens 2019), or in the main survey as in PISA 2012 (OECD 2014).

With paper-based questionnaires, there is only very limited information on response behavior to assert the quality of the instruments. In student sessions, report forms completed by test administrators provide information regarding certain aspects, such as timing or anomalies and deviations from uniform conditions during assessment sessions. Therefore, information on the way in which respondents react to the questionnaire material delivered on paper is generally only obtained through pretesting at the pilot and field trial stages, which generates narrative and partly anecdotal information.

Electronic delivery of instruments, however, provides information beyond this, and allows statistical and other analyses of the substantial response data in conjunction with log data. Kroehne and Goldhammer (2018) proposed a framework to conceptualize, represent, and use data from technology-based assessment, explicitly including log data collected for contextual questionnaires. Their model encompasses an access-related category (including, e.g., assessment setting and device information), a response-related category (e.g., input-related events), finally a process category (e.g., navigation). Hu's (2016) cross-cultural survey guidelines provide a similar conceptualization and recommendations for the purpose of reviewing and

explaining non-response at the case and item level, or analyzing aberrant responses and/or “satisficing” (a term combining satisfy and suffice, referring to the idea that people do not put as much effort into responding as they should; see Tourangeau et al. 2000).

A particular benefit from the approach proposed by Kroehne and Goldhammer (2018) relates to the aim of generating indicators from individual events, states, and the in-between transitions that have the potential of informing survey designers about response processes from the perspective of the respondents (e.g., regarding timing, navigation, drop-out and non-response), and the individual questions and items (such as average time needed to respond by assessment language, scrolling, or changes of responses). Data and indicators can then generate insights with respect to the technical behavior of systems, access limitations, and device preferences, which may all assist with optimization at the system level. More importantly, the data and indicators can provide powerful insights into the functioning of the question materials, identifying questions and items requiring a disproportionate long response time or indicating error-prone recollection requirements (Tourangeau et al. 2000). In addition, process data may provide data about respondents’ engagement, or disengagement, and the extent to which the collected information validly relates to the questions or is a result of inattentive or otherwise aberrant behavior, which may include deliberate falsification (for an overview of detection methods, see, e.g., Steedle et al. 2019).

The second promising aspect of electronic delivery relates to the ability to deliver questionnaires in a non-linear way to respondents. To date, virtually all ILSAs deliver one version of a questionnaire to respondents at the main data collection stage. However, there is a strong interest in broadening the conceptual depth and breadth of measures in questionnaires to yield additional insights for educational policy. The situation is similar to one of the most important late 20th century advancements in cognitive assessments in education, the so-called “new design” implemented by the National Assessment of Educational Progress (NAEP) to broaden the assessment’s domain scope and insights for policy while managing the response burden for individuals (Mislevy et al. 1992). Essentially, the responses (and, by extension, the derived variables) for items not administered to an individual student were treated as a missing data problem, addressed through the use of IRT and latent regression modeling based on Bayesian approaches accounting for imputation variance.

IEA’s technical standards (Martin et al. 1999) acknowledged the similar potential for questionnaires early on:

Consider whether matrix sampling may be appropriate in the development of the questionnaire. Matrix sampling in this case means that not all respondents are asked all questions. Although this method will add some cost and complexity, it can greatly reduce response burden. Matrix sampling should only be considered if the study objectives can be met with adequate precision.

While this matrix sampling is nowadays firmly established in internationally comparative cognitive assessments, it remains to be seen if such an approach can be carried over to the case of questionnaires. Some relevant research related to these aspects has already been undertaken (Adams et al. 2013; Kaplan and Su 2018; von Davier 2014). Electronic delivery coupled with modern statistical approaches,

such as predictive mean matching, is believed to have potential for more elaborate sequencing of materials, matrix sampling approaches, and other aspects. Insights from such research could in due time become the basis for new technical standards for future IEA studies.

5.7 Conclusions

Questionnaires are a well-established component of ILSAs and provide crucial information on context variables and potential outcome variables. From earlier uses as auxiliary instruments to provide data to explain achievement results, questionnaires have grown in importance in recent international studies.

The sophistication of development and implementation procedures has also grown in recent decades. With regard to cross-national educational research, there has been an increasing recognition of the importance of considering the validity and reliability of questionnaire instruments and measures. The requirement of cross-national comparability is a crucial element that is emphasized by the fact that, across ILSAs, increasing attention is paid to questionnaire outcomes as a way of comparing student learning and educational contexts. The recognition of the potential bias resulting from differences in national contexts when using questionnaires in these studies has led to an increased focus on thorough review, piloting, and trialing of material as part of questionnaire development, with further analyses aimed at detecting or ameliorating non-equivalence of translations or resulting measures.

Here, we have described some of the main approaches to the questionnaires applied in international studies of educational achievement, their targeting and tailoring to distinct respondent groups, the variety of measures, indicators, and formats, the procedures typically implemented to ensure thorough development, challenges resulting from cross-national measurement, and the transition from paper-based to electronic delivery. In particular the challenges of maximizing measurement invariance across highly diverse national contexts and the opportunities provided by computer-based delivery are expected to result in interesting developments in the near future, which may lead to further changes and improvements in the approach to questionnaire elaboration and implementation in ILSAs.

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, 5. <https://largescaleassessmentsineducation.springeropen.com/articles/10.1186/2196-0739-1-5>.
- Agrusti, G., Ainley, J., Losito, B., & Schulz, W. (2018). ICCS questionnaire development. In W. Schulz, R. Carstens, B. Losito, & J. Fraillon (Eds.), *ICCS 2016 technical report* (pp. 21–32). Amsterdam, The Netherlands: IEA. <https://www.iea.nl/publications/technical-reports/iccs-2016-technical-report>.

- Ainley, J., & Carstens, R. (2018). *Teaching and Learning International Survey (TALIS) 2018 conceptual framework*. OECD Education Working Papers, No. 187. Paris, France: OECD Publishing. <https://doi.org/10.1787/799337c2-en>.
- Ainley, J., & Schulz, W. (2020). ICILS 2018 questionnaire development. In J. Fraillon, J. Ainley, W. Schulz, T. Friedman, & S. Meyer (Eds.), *ICILS 2018 technical report* (pp. 39–47). Amsterdam, the Netherlands: IEA. <https://www.iea.nl/publications/technical-reports/icils-2018-technical-report>.
- Byrne, B. M., & van de Vijver, F. J. R. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, 10(2), 107–132.
- Carstens, R. (2019). Development of the teacher and principal questionnaires. In OECD (Ed.), *TALIS 2018 technical report* (pp. 53–84). Paris, France: OECD Publishing.
- Carstens, R., & Pelgrum, J. P. (2009). *Second Information Technology in Education Study. SITES 2006 technical report*. Amsterdam, The Netherlands: IEA. <https://www.iea.nl/publications/technical-reports/second-information-technology-education-study-technical-report>.
- Couper, M. (2008). *Designing effective web surveys*. Cambridge, UK: Cambridge University Press. <https://doi.org/10.1017/CBO9780511499371>.
- Crowne, D. P., & Marlowe, D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, 24(4), 349–354.
- Desa, D., van de Vijver, F., Carstens, R., & Schulz, W. (2018). Measurement invariance in international large-scale assessments: Integrating theory and method. In T. P. Johnson, B.-E. Pennell, I. Stoop, & B. Dorner (Eds.), *Advances in comparative survey methodology* (pp. 881–910). Chichester, UK: Wiley.
- Dillman, D. A., Tortora, R. D., & Bowker, D. (1998). *Principles for constructing web surveys: An initial statement*. Technical report No. 98-50. Pullman, WA: Washington State University Social and Economic Sciences Research Center.
- European Commission/EACEA/Eurydice. (2017). *Citizenship education at school in Europe–2017*. Eurydice Report. Luxembourg: Publications Office of the European Union. https://eacea.ec.europa.eu/national-policies/eurydice/content/citizenship-education-school-europe-%E2%80%93-2017_en.
- Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2020). *Preparing for life in a digital world: IEA International Computer and Information Literacy Study 2018 international report*. Cham, Switzerland: Springer. <https://www.iea.nl/publications/study-reports/preparing-life-digital-world>.
- Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Gebhardt, E. (2014). *Preparing for life in a digital age. The IEA International Computer and Literacy Information Study international report*. Cham, Switzerland: Springer. <https://www.iea.nl/publications/study-reports/international-reports-iea-studies/preparing-life-digital-age>.
- Fraser, B. J. (1982). Differences between student and teacher perceptions of actual and preferred classroom learning environment. *Educational Evaluation and Policy Analysis*, 4(4), 511–519.
- Harkness, J., Bilgen, I., Córdova Cazar, A., Hu, M., Huang, L., Lee, S., Liu, M., Miller, D., Stange, M., Villar, A., & Yan, T. (2016). *Questionnaire design. Guidelines for best practice in cross-cultural surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. <http://www.ccsr.isr.umich.edu/>.
- Heine, S. J., Lehman, D. R., Peng, K., & Greenholtz, J. (2002). What's wrong with cross-cultural comparisons of subjective Likert scales? The reference group effect. *Journal of Personality and Social Psychology*, 82(6), 903–918.
- Hooper, M. (2016). Developing the TIMSS 2015 context questionnaires. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 2.1–2.8). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. <http://timss.bc.edu/publications/timss/2015-methods/chapter-2.html>.

- Hooper, M., & Fishbein, B. (2017). Developing the PIRLS 2016 context questionnaires. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 2.1–2.8). Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. <https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-2.html>.
- Hu, M. (2016). *Paradata and other auxiliary data. Guidelines for best practice in cross-cultural surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. <http://www.ccsr.isr.umich.edu/>.
- IEA. (2020a). IEA studies [webpage]. Amsterdam, The Netherlands: IEA. <https://www.iea.nl/studies/ieastudies>.
- IEA. (2020b). Publications [webpage]. Amsterdam, The Netherlands: IEA. <https://www.iea.nl/publications>.
- Johnson, T. P., & van de Vijver, F. J. R. (2003). Social desirability in cross-cultural research. In J. A. Harkness, F. J. R. van de Vijver, & P. P. Mohler (Eds.), *Cross-cultural survey methods* (pp. 195–204). Hoboken, New Jersey: Wiley.
- Johnson, T. P., Pennell, B. E., Stoop, I., & Dorier, B. (Eds.). (2018). *Advances in comparative survey methodology*. Chichester, UK: John Wiley & Sons Ltd.
- Jude, N., & Kuger, S. (2018). *Questionnaire development and design for international large-scale assessments (ILSAs): Current practice, challenges, and recommendations*. Workshop series on methods and policy uses of international large-scale assessments (ILSAs). Washington, DC: National Academy of Education. <http://naeducation.org/wp-content/uploads/2018/04/Jude-and-Kuger-2018-FINAL.pdf>.
- Kaplan, D., & Su, D. (2018). On imputation for planned missing data in context questionnaires using plausible values: a comparison of three designs. *Large-scale Assessments in Education*, 6, 6. <https://largescaleassessmentsineducation.springeropen.com/articles/10.1186/s40536-018-0059-9>.
- Kerr, D., Schulz, W., & Fraillon, J. (2011). The development of regional instruments. In W. Schulz, J. Ainley, & J. Fraillon (Eds.), *ICCS 2009 technical report* (pp. 45–49). Amsterdam, The Netherlands: IEA. <https://www.iea.nl/publications/technical-reports/iccs-2009-technical-report>.
- Kerr, D., Sturman, L., Schulz, W., & Bethan, B. (2010). *ICCS 2009 European report. Civic knowledge, attitudes and engagement among lower secondary school students in twenty-four European countries*. Amsterdam, The Netherlands: IEA. <https://www.iea.nl/publications/study-reports/international-reports-iea-studies/iccs-2009-european-report>.
- Kroehne, U., & Goldhammer, F. (2018). How to conceptualize, represent, and analyze log data from technology-based assessments? A generic framework and an application to questionnaire items. *Behaviormetrika*, 45, 527. <https://doi.org/10.1007/s41237-018-0063-y>.
- Kuger, S., Klieme, E., Jude, N., & Kaplan, D. (Eds.). (2016). *Assessing contexts of learning: An international perspective*. Cham, Switzerland: Springer.
- Law, N., Pelgrum, W. J., & Plomp, T. (2008). *Pedagogy and ICT use in schools around the world. Findings from the IEA SITES 2006 study*. Hong Kong: CERC-Springer.
- Lietz, P. (2010). Research into questionnaire design: A summary of the literature. *International Journal of Market Research*, 52(2), 249–272. <https://doi.org/10.2501/S147078530920120X>.
- Lietz, P. (2017). Design, development and implementation of contextual questionnaires in large-scale assessments. In P. Lietz, J. Cresswell, K. Rust, & R. Adams (Eds.), *Implementation of large scale education assessments* (pp. 92–136). Chichester, UK: Wiley.
- Losito, B., Agrusti, G., Damiani, V., & Schulz, W. (2018). *Young people's perceptions of Europe in a time of change: IEA International Civic and Citizenship Education Study 2016 European report*. Amsterdam, The Netherlands: IEA. <https://www.iea.nl/publications/study-reports/international-reports-iea-studies/young-peoples-perceptions-europe-time>.
- Martin, M. O. Rust, K., & Adams, R. J. (Eds.). (1999) *Technical standards for IEA studies*. Amsterdam, The Netherlands: IEA. <https://www.iea.nl/publications/iea-reference/technical-standards-iea-studies>.
- McCrae, R. R., & Costa, P. T. (1983). Social desirability scales: More substance than style. *Journal of Consulting and Clinical Psychology*, 51, 882–888.

- Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, 29(2), 133–161. <https://doi.org/10.2307/1434599>.
- Mullis, I. V. S., & Martin, M. O. (Eds.). (2017). *TIMSS 2019 assessment frameworks*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College. <https://www.iea.nl/publications/assessment-framework/timss-2019-assessment-frameworks>.
- OECD. (2014). *PISA 2012 technical report*. Paris, France: OECD Publishing. <https://www.oecd.org/pisa/pisaproducts/PISA-2012-technical-report-final.pdf>.
- OECD. (2019a). TALIS: The OECD Teaching and Learning International Survey. Paris, France: OECD. <http://www.oecd.org/education/talis/>.
- OECD. (2019b). OECD Starting Strong Teaching and Learning International Survey. Paris, France: OECD. <http://www.oecd.org/education/school/oecd-starting-strong-teaching-and-learning-international-survey.htm>.
- Presser, P., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., et al. (Eds.). (2004). *Methods for testing and evaluating survey questionnaires*. Hoboken, NJ: Wiley.
- Schulz, W. (2009). Questionnaire construct validation in the International Civic and Citizenship Education Study. In M. von Davier & D. Hastedt (Eds.), *Issues and methodologies in large-scale assessments*, IERI Monograph Series (Vol. 2, pp. 113–135). Hamburg, Germany: IERI. http://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_05.pdf.
- Schulz, W., & Fraillon, J. (2011). The analysis of measurement equivalence in international studies using the Rasch model. *Educational Research and Evaluation*, 17(6), 447–464.
- Schulz, W., & Nikolova, R. (2004). Translation procedures, field operations and quality assurance. In W. Schulz & H. Sibbern (Eds.), *IEA Civic Education Study technical report* (pp. 27–40). Amsterdam, The Netherlands: IEA. <https://www.iea.nl/publications/technical-reports/iea-civic-education-study-technical-report>.
- Schulz, W., Ainley, J., Cox, C., & Friedman, T. (2018a). *Young people's views of government, peaceful coexistence, and diversity in five Latin American countries. The International Civic and Citizenship Education Study 2016 Latin American Report*. Amsterdam, The Netherlands: IEA. <https://www.iea.nl/publications/study-reports/international-reports-iaa-studies/young-peoples-views-government>.
- Schulz, W., Ainley, J., Fraillon, J., Kerr, D., & Losito, B. (2010). *ICCS 2009 international report. Civic knowledge, attitudes and engagement among lower secondary school students in thirty-eight countries*. Amsterdam, The Netherlands: IEA. <https://www.iea.nl/publications/study-reports/international-reports-iea-studies/iccs-2009-international-report>.
- Schulz, W., Ainley, J., Friedman, T. & Lietz, P. (2011). *ICCS 2009 Latin American report: Civic knowledge and attitudes among lower secondary students in six Latin American countries*. Amsterdam, The Netherlands: IEA. <https://www.iea.nl/publications/study-reports/international-reports-iea-studies/iccs-2009-latin-american-report>.
- Schulz, W., Ainley, J., Fraillon, J., Losito, B., & Agrusti, G. (2016). *IEA International Civic and Citizenship Education Study 2016 assessment framework*. Cham, Switzerland: Springer. <https://www.iea.nl/publications/assessment-framework/iea-international-civic-and-citizenship-education-study-2016>.
- Schulz, W., Ainley, J., Fraillon, J., Losito, B., Agrusti, G., & Friedman, T. (2018b). *Becoming citizens in a changing world: The International Civic and Citizenship Education Study 2016 international report*. Cham, Switzerland: Springer. <https://www.iea.nl/publications/study-reports/international-reports-iaa-studies/becoming-citizens-changing-world>.
- Steedle, J. T., Hong, M., & Cheng, Y. (2019). The effects of inattentive responding on construct validity evidence when measuring social-emotional learning competencies. *Educational Measurement: Issues and Practice*, 38, 101–111. <https://doi.org/10.1111/emip.12256>.
- Survey Research Center. (2016). *Guidelines for best practice in cross-cultural surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. <http://www.ccsr.isr.umich.edu/>.

- Torney-Purta, J., Lehmann, R., Oswald, H. & Schulz, W. (2001). *Citizenship and education in twenty-eight countries: Civic knowledge and engagement at age fourteen*. Amsterdam, The Netherlands: IEA. <https://www.iea.nl/publications/study-reports/international-reports-iea-studies/citizenship-and-education-twenty-eight>.
- Torney, J., Oppenheim, A. N., & Farnen, R. F. (1975). *Civic education in ten countries: An empirical study*. New York, NY, USA: Wiley.
- Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, UK: Cambridge University Press. <https://doi.org/10.1017/CBO9780511819322.004>.
- von Davier, M. (2014). Imputing proficiency data under planned missingness in population models. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis* (pp. 175–201). Boca Raton, FL: CRC Press.
- Van de Gaer, E., Grisay, A., Schulz, W., & Gebhardt, E. (2012). The reference group effect: An explanation of the paradoxical relationship between academic achievement and self-confidence across countries. *Journal of Cross-Cultural Psychology*, 43, 1205–1228.
- Van de Vijver, F. J. R., & He, J. (2014). *Report on social desirability, midpoint and extreme responding in TALIS 2013*. OECD Education Working Papers, No. 107. Paris, France: OECD Publishing. https://www.oecd-ilibrary.org/education/report-on-social-desirability-midpoint-and-extreme-responding-in-talis-2013_5jxswcfwt76h-en.
- Van de Vijver, F. J. R., Jude, N., & Kuger, N. (2019). Challenges in international large-scale educational surveys. In B. Denman, L. E. Suter, & E. Smith (Eds.), *Sage handbook of international comparative research* (pp. 83–102). London, UK: SAGE.

Wolfram Schulz is a Principal Research Fellow (formerly Research Director International Surveys) at the Australian Council for Educational Research (ACER) where he has worked on a large number of national and international large-scale assessment studies. He is International Study Director of the IEA International Civic and Citizenship Education Study (ICCS) and Assessment Coordinator for the IEA International Computer and Information Literacy Study (ICILS). He is also a member of the IEA Technical Executive Group (TEG).

Ralph Carstens is a Senior Research Advisor at the International Association for the Evaluation of Educational Achievement (IEA) in Hamburg. With a professional background in primary and lower secondary teaching, Ralph acquired deep knowledge and experience relating to the conceptual development, instrument design, survey methodology, data work, analysis, reporting, and communication of international large-scale surveys in education. Ralph recently co-directed the IEA International Civic and Citizenship Education Study (ICCS) 2016 study and chaired the Questionnaire Expert Group for the OECD Teaching and Learning International Survey (TALIS). He coordinates the initiation and conceptual development of new studies, projects, and partnerships with a focus on non-IEA studies. He further provides advice on technical and methodological solutions and innovations to IEA management.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.