Chapter 4

Assessment Content Development

Liz Twist and Julian Fraillon

Abstract The very nature of international large-scale assessments (ILSAs) raises unique challenges for test developers. The surveys are used across many different countries and translated into multiple languages; they must serve as sound measurement instruments across a very wide range of attainment whilst also engaging students from diverse backgrounds and with very different educational experiences. In order for the surveys to meet the stated aims to evaluate, understand, and improve education worldwide, policymakers must have confidence that the conclusions drawn from the results are relevant and meaningful in their particular jurisdictions. The International Association for the Evaluation of Educational Achievement (IEA) has a comprehensive assessment development process; assessment experts and representatives of participating countries review the material at key points during the test development and throughout strict standards for quality assurance are applied. Care is taken to ensure that consideration is given to diverse perspectives whilst maintaining the integrity and rigor of the assessment instruments as specified in the assessment framework. This established and thorough approach has enabled the assessments to evolve whilst continuing to meet the diverse needs of stakeholders.

Keywords Assessment development • Comparative studies • International Civic and Citizenship Education Study (ICCS) • International Computer and Information Literacy Study (ICILS) • International large-scale assessments (ILSAs) • Progress in International Reading Literacy Study (PIRLS) • Trends in International Mathematics and Science Study (TIMSS)

L. Twist (✉)

National Foundation for Educational Research (NFER), The Mere, Upton Park, Slough, Berkshire SL1 2DQ, UK

e-mail: l.twist@nfer.ac.uk

J. Fraillon

Australian Council for Educational Research (ACER), 19 Prospect Hill Rd, Camberwell, VIC 3124, Australia

e-mail: Julian.Fraillon@acer.org

4.1 Introduction

IEA is faced with considerable challenges when developing appropriate questions for international large-scale assessments (ILSAs) that are to be used across a broad range of countries, languages, and cultures. For the assessments to contribute to meeting IEA's stated aim, to evaluate, understand, and improve education worldwide, the assessments must have robust technical functioning and yet retain credibility with stakeholders. Whilst there exist some internationally recognized technical standards relating to the development of educational assessments (see Sect. 4.5), even within countries, there is unlikely to be unanimous agreement about what constitutes an appropriate and high quality assessment instrument. This diversity of views is magnified when considering the number and variety of countries that participate in IEA surveys and the challenges associated with putting into practice the principles and standards relating to educational assessment.

In this chapter, we aim to show how the technical quality and strength of IEA assessments is a result of deliberate strategies to maximize the benefits of the diverse perspectives of IEA's researchers, stakeholders, and expert consultants, and how a collaborative and consultative approach leads to the development of high quality measurement instruments. We discuss the process of assessment item development for IEA surveys, looking at the range of item types and the role of participating countries in item development. We consider the approaches adopted to ensure quality in item development.

4.2 Key Features of ILSAs that Influence Assessment Content Development

ILSAs are, by definition, not aligned to any specific country's curriculum or framework. Further to this, the specificity and content of curricula, standards, or frameworks vary greatly across countries and according to learning areas being assessed. In the IEA context,¹ reading, assessed in the Progress in International Reading Literacy Study (PIRLS), and mathematics and science, assessed in the Trends in International Mathematics and Science Study (TIMSS), are core learning areas with consequently strong and explicit representation in country curricula and, where applicable, local or national assessments within countries. In contrast, there is far greater variation across countries in the explicitness and emphasis given in the curriculum to civics and citizenship education, computer and information literacy (CIL), and computational thinking (CT), skills that are measured in the International Civic and Citizenship Education Study (ICCS) and the International Computer and

---

¹For more than 60 years, IEA has been a leading figure in the field of comparative studies of education. Their activities cover all aspects of educational research. Current international IEA studies include TIMSS, PIRLS, ICCS, and ICILS (see IEA 2020).

Information Literacy Study (ICILS). While in all ILSAs it is essential for the assessment content to be drawn from a broad interpretation of the construct defined in the assessment framework (see Chap. 3 for further details of assessment framework development), the different approaches and level of curriculum detail across countries introduce a unique set of challenges to the development of test content in ILSAs.

In order to maximize curricula/domain coverage, a matrix survey design is typically used, in which individual students complete a sub-sample of items, the entirety of which assess the defined domain. This requires the development of an extensive pool of items, each of which is suitable for use in each of the participating countries. This pool needs to include items with a very wide range of difficulty in order to provide all participating countries with sufficient precision in the outcomes of the surveys to meet their objectives. This is particularly challenging when there is a wide range of student achievement across countries.

There is a significant challenge of developing a pool of items that is suitable for use across a range of countries, and agreed by country representatives to represent the learning area as it is understood and assessed within each country. While the test content is developed with reference to a common assessment framework (rather than to any given country's curriculum), expert judgements of the suitability of each item to assess the specified content in each country rightly take into account existing relevant assessments in the same or similar learning areas within countries. In learning areas such as reading, mathematics, and science that are assessed in PIRLS and TIMSS, many countries have well-established pools of existing items that are used in national or local assessments which can provide a frame of reference. ILSA assessment content, while governed by the specifications of the assessment framework, must also be recognizably relevant to the assessment of learning areas as they are understood and represented in national assessment contexts. Evaluating the coherence between ILSA assessment content and national assessment contexts can be more difficult in studies such as ICCS and ICILS. In such studies, while some participating countries may have explicit curricula and standards that, together with contributions from relevant academic literature and expert judgements, can contribute to the content of the assessment framework, many countries may not have existing pools of assessment items that national experts can refer to when evaluating the suitability of the ILSA test items for use in their national contexts. In these cases, expert judgements of the appropriateness of the assessment content may need to be based on more abstract conceptualizations of what is likely to be relevant and suitable rather than in comparison with what is known already to work within countries.

In addition to the objective of measuring and reporting student achievement at a given point in time, a key reason that many countries choose to participate in ILSAs is to monitor achievement over time. ILSAs have varying cycles of data collection. In IEA studies, the PIRLS cycle is five years, TIMSS is four years, ICCS is seven years, and ICILS is five years. This requirement for longevity is discussed further in Chap. 2, but it does place an additional demand on test item development. That is, the item pool needs to include items that are likely to be suitable for use in future cycle(s), as well as in the cycle in which they are developed. Items that are used in more than two cycles in one of the listed ILSAs may therefore need to be appropriate over a

period spanning more than 14 years. In all learning areas this can pose challenges for item development. In IEA studies, this clearly poses challenges for ICILS when working in the domain of rapidly evolving technologies, but similar challenges are emerging as all studies transition to computer-based delivery.

4.3 Validity in International Large-Scale Assessments

The assessment review processes described in this chapter are operational manifestations of the aim to maximize the validity of the assessments. In this context, validating the assessment requires an evaluation of the evidence used to support particular interpretations of the survey results. While this includes an evaluation of the assessment content that is the focus of this chapter, the frame of reference for the review of the validity of the assessment is broader than the contents of the assessment itself (Kane 2013).

The conceptualization of validity proposed by Kane (2013) requires that validity focuses on the use of the assessment outcomes. Also relevant is the work of Oliveri et al. (2018, p. 1), who proposed a conceptual framework to assist participating countries to:

- Systematically consider their educational goals and the degree to which ILSA participation can reasonably help countries monitor progress toward them;
- Use an argument model to analyze claims by ILSA programs against the background of a country's specific context; and
- More clearly understand intended and unintended consequences of ILSA participation.

Others, such as Stobart (2009), have recognized the complexity of producing a validity argument when assessments may be used for multiple and varied purposes. His reservations concerned the validity in the use of one country's national assessments to which many purposes had become attached; the demand is increasingly complex when it concerns the use of an assessment in dozens of countries.

In this chapter, we elaborate on the process of item development in IEA surveys. This is the foundation for the two key sources of validity evidence: expert review, and item and test analysis of the survey data. Other sources of validity evidence may be available within countries; for example the association between performance in the surveys and in other national assessments, but inevitably this is local evidence. While this chapter focuses on how the assessment instrument development process is used to evaluate the validity of the instrument, this notion of validity works in the larger framework (suggested by Kane 2013) in which the evaluation relates to the suitability of the instruments to elicit data that can be used to support defensible interpretations relating to student outcomes in the areas of learning being researched.

4.4 The Assessment Frameworks

As explored in Chap. 3, the assessment frameworks define the construct/s to be assessed and the nature of the assessment to be developed. These documents, publicly available and rooted in the research theory and evidence, are reviewed and revised by expert groups in the early stages of each cycle. They are used to guide the content development but also have the potential to support participation decisions and appropriate interpretation of the outcomes.

In their definitions of the constructs to be assessed, the assessment frameworks inevitably shape the assessment design. In the case of reading, for example, the PIRLS framework defines reading literacy as follows:

Reading literacy is the ability to understand and use those written language forms required by society and/or valued by the individual. Readers can construct meaning from texts in a variety of forms. They read to learn, to participate in communities of readers in school and everyday life, and for enjoyment (Mullis and Martin 2019).

This focus on reading as a meaning-making process requires the assessment to ensure that participating students engage with and respond to the written texts. This response takes a written form. There is no element of the PIRLS assessment that specifically assesses decoding, namely students' ability to convert graphemic (or logographic) forms into sounds. Whilst decoding is implicit in all reading, the starting point for the PIRLS assessment materials is the individual and, in most cases, silent reading of written texts ("passages") and the assessment is of students' ability to comprehend them by answering, in writing, the written questions.

In ICILS, CIL is defined as:

...an individual's ability to use computers to investigate, create, and communicate in order to participate effectively at home, at school, in the workplace and in society (Fraillon et al. 2019, p. 18).

and CT is defined as:

...an individual's ability to recognize aspects of real-world problems which are appropriate for computational formulation and to evaluate and develop algorithmic solutions to those problems so that the solutions could be operationalized with a computer (Fraillon et al. 2019, p. 27).

In each of the ICILS constructs, there is a clear emphasis on the use of computers as problem-solving tools. For CIL there is an emphasis on information gathering and communication, whereas for CT the emphasis is on conceptualizing and operationalizing computer-based solutions to problems. Both definitions suggest the use of computer delivered instruments and an emphasis on achievement being measured and demonstrated in real-world contexts. In response to these demands, the ICILS CIL and CT instruments consist of modules comprising sequences of tasks linked by a common real-world narrative theme per module (see Fraillon et al. 2020).

It is particularly important that a clear exposition of the knowledge, skills, and understanding being assessed in the international surveys is provided. While this

begins with the assessment framework, any review of the assessment instruments includes consideration of the degree to which these instruments address the assessment outcomes articulated by the framework. As part of the development process, each assessment item is mapped to the relevant framework and the accuracy and defensibility of these mappings is one aspect of the validity review. While it is not possible to collect other validity evidence, such as the relationship between performance on the survey and performance on another assessment in broadly the same domain (concurrent validity) on an international level, it is possible that some countries could potentially undertake such an exercise.

4.5 Stimulus Material and Item Development: Quality Criteria Associated with Validity

There are well-established criteria that all assessment material should be evaluated against. The *Standards for Educational and Psychological Testing* (American Educational Research Association et al. 2014), for example, use the concepts of validity, reliability, and fairness as organizing perspectives from which to evaluate the quality of assessment material. For the purpose of this chapter, we will address eight groups of evaluation criteria that are routinely implemented in the development of assessment materials in IEA studies:

- representation of the construct
- technical quality
- level of challenge
- absence of bias
- language and accessibility
- cultural and religious contexts
- engagement of test-takers
- scoring reliability.

These quality criteria are applied during all phases of the materials development process.

4.5.1 Representation of the Construct

As described in the previous section, the assessment constructs in ILSAs are defined and explicated in detail in an assessment framework (see Chap. 3 for further details). In the context of ILSA, the assessment constructs do not represent any given national curriculum but are designed to be relevant to and recognizable within national curricula. Once an ILSA assessment construct has been accepted by countries, it is essential that the assessment instrument provides a true representation of the

construct. This evaluation, conducted by assessment developers, country representatives, and other experts takes place throughout the materials development process. When considering this criterion, it is essential that reviewers consider the construct being used in the study without conflating it with what might be used in local (national) contexts. For example, in PIRLS, poetry is not included in the assessment because of the specific challenges associated with translation. In many curricula, the reading and comprehension of poetry is mandated. Despite this omission, what is included in PIRLS is a wide representation of reading comprehension and is accepted as such by stakeholders.

All assessments in IEA studies are also carefully mapped to their constructs. The assessment frameworks typically specify the proportions of items addressing different aspects of the constructs to ensure a full and appropriate reflection of the constructs in the instruments. As part of the review process, both the accuracy of the mapping of items to the constructs and the degree to which the total instrument meets the design specifications in the framework are considered.

4.5.2 Technical Quality

While the technical quality of test materials could be considered a property of the representation of the construct in the items, it warrants independent explication, as it is central to the materials review. Questions that need to be considered in the technical review of the materials can include, but are not limited to:

- Is the material clear, coherent, and unambiguous?
- Is the material self-contained? Or does it assume other prior knowledge, and, if so, is this appropriate?
- Are there any “tricks” in materials that should be removed?
- Is each key (the correct answer to a multiple choice question) indisputably correct?
- Are the distractors (the incorrect options to a multiple choice question) plausible but indisputably incorrect?
- Do the questions relate to essential aspects of the construct or do they focus on trivial side issues?
- Is the proposed item format the most suitable for the content in each case?
- Are there different approaches to arriving at the same answer? If so, do these different approaches represent equivalent or different levels of student ability and should this be reflected in the scoring?
- Is there any local dependence across items within a unit (testlet) or across the instrument? Local dependence occurs when either the content of or the process of answering one question affects the likelihood of success on another item.

4.5.3 *Level of Challenge*

For the assessments to function well in psychometric terms across the participating countries, they must adequately measure the skills of the highest and lowest attainers. Within all countries there is a range of attainment in the assessed domains; the difference between countries is often in the proportions of students at different points across the range. Of course, there are some notably high achieving countries: in TIMSS 2015 at grade 4, for example, 50% of students in Singapore reached the advanced benchmark. In contrast, in 12 countries, fewer than three percent of students reached the advanced benchmark and, in four of these countries, fewer than 50% of students reached the low benchmark (Mullis et al. 2016). As a measurement exercise, students need to complete some items that they find straightforward and some that are challenging: if a student scores maximum points the measure is not providing full information about their capabilities; similarly a failure to score any points is not informative about what skills have been developed. When selecting content for the main survey data collection, careful consideration is given to the relative proportion and allocation of items of varying difficulty across the assessment. More recently, some studies have included plans for an approach that allows for the balance of item difficulties to vary across countries to better match known profiles of achievement within countries.

4.5.4 *Absence of Bias*

Whilst this aspect of the quality criteria is part of the psychometric analysis, it is also a consideration during the development process. Bias occurs when factors other than those identified as integral to the assessment impact on the scores obtained by students, meaning that students with the same underlying ability do not achieve equivalent scores. In technical terms, this is construct irrelevant variance. There are a number of potential sources of bias that test developers are mindful of. The benefit of prior experience can be evident in performance on a reading test, for example, where, independent of reading ability, the assessment rewards some test-takers for existing knowledge. A reading assessment would not directly include content from a reading program widely used in a participating country but it would consider for inclusion material published in particular countries as this provides a necessary level of authenticity. The review for bias is one in which the perspectives of country representatives is crucial, as they may identify particular content, themes, or topics that may unfairly advantage or disadvantage test-takers in their national context. The psychometric measures of bias that inform assessment development rely on there being sufficient numbers of test-takers in the sub-groups for comparison. Routinely in ILSA, bias across countries is measured at the item level in what is referred to as “item-by-country interaction” and bias within countries (and cross-nationally) is measured between female and male test-takers.

4.5.5 *Language and Accessibility*

The assessment content is developed and reviewed in English. In all assessment development there is a need for consideration of the precision in language alongside the amount of reading required. In the case of international surveys, a further consideration is the impact of translation, as discussed in Chap. 6. During the development phase the onus is on country representatives to be alert to any particular concerns about the feasibility of translating particular words and phrases. It is frequently the case that, in discussion, often within multilingual groups, alternative words and phrases are identified that function equally well within the assessment.

Reading ability should not influence test performance when reading is not the construct being assessed. For this reason, accommodations such as readers may be used in the administration of an assessment of science, for example. In studies such as TIMSS, ICCS, and ICILS, where reading is not the domain being assessed, there is a clear and deliberate effort to keep the reading load to a minimum. A rule-of-thumb is that the typical reading load in a non-reading assessment should be equivalent to the level attained by students that are roughly two grades below the grade level of the students being tested. The reading load is primarily influenced by sentence length, sentence structure, and vocabulary use. In some cases, it is feasible to use readability indexes to support the evaluation of the reading load of materials, however, interpreting the output of a readability index must be done with careful consideration of the domain being assessed. For example, in ICCS, terms such as *democracy* or *sustainable development* may represent essential content of the domain, but also inflate the reading load of materials when measured using a readability index. In addition, as the original materials are developed in English, readability index outcomes applied to English language text cannot be assumed to be appropriate when considering how the text will appear under translation to languages other than English.

4.5.6 *Cultural and Religious Contexts*

Developers of all assessments used within a single country need to be alert to potential cultural or religious issues that may impact on how the assessment is interpreted. It is an even greater focus when the assessments are deployed internationally. The issues are different according to the domains being assessed. For example, certain concepts are accepted as legitimately part of a science assessment and, in fact, required in order to ensure as comprehensive an assessment as possible, but may not be as readily accepted in a reading assessment. Similarly, some texts, such as traditional fables, may focus on explanations of natural phenomena that would have no place in a science assessment and may also challenge some beliefs, yet they are a part of what is accepted as the broad literary canon that may legitimately be included in a reading assessment.

This criterion includes consideration not just of the items but also the images and contexts incorporated into some assessments. When selecting contexts for ICILS content there are, for example, varying rules and laws across countries relating to grade 8 students' access to and engagement with social media platforms. The involvement of representatives of participating countries in the development process ensures that many perspectives are considered at an early stage.

4.5.7 Engagement of Test-Takers

There is ample evidence that more engaged students perform better. This is evident in better learning in the classroom (e.g., Saeed and Zyngier 2012) or in test-taking (e.g., Penk et al. 2014). While test developers make no attempt to entertain students, they do aspire to present engaging material to the young people completing the assessment, namely the sort of content that Ryan and Deci (2000) described as being “intrinsically interesting” for the majority. This may be in the contexts selected for scenarios for tasks, or in the texts selected. In PIRLS, for example, students are asked to indicate how much they enjoyed reading specific texts at the field trial stage. This is one of the sources of evidence that is considered when selecting the final content.

In addition, the assessment should have a degree of coherence for the student, even though each student completes only a part of the overall assessment. Each student will be exposed to items that assess more than solely number operations; for example, in TIMSS and in ICCS, each test booklet includes items that assess content associated with all four content domains in approximately the same proportions as those specified in the assessment framework for the test instrument as a whole.

4.5.8 Scoring Reliability

At the heart of reliable scoring is consistency. The scoring guides must be interpreted in the same way, ensuring that the same responses achieve the same score, regardless of who the scorer is. Scoring reliability is measured across countries in each survey and within countries by looking at consistency between cycles (part of the trend measure). To establish this consistency, those who undertake the scoring require facility in both the language(s) of the test in their country and also in English. In IEA studies, the international scorer training (i.e., the training of the people responsible for scoring and scorer training within each participating country) is conducted in English, and the scoring guides and scoring resources are presented in English. Some countries choose to translate the scoring materials to the language of the national test, and to run their national scoring in their language of testing; in PIRLS 2016, over half participating countries supplemented their scoring materials with example responses produced within their particular country (Johansone 2017).

Good assessment development practice requires the scoring guides to be developed alongside the items; developers need to document the answers that they are expecting to receive credit and to use these to confirm the process being assessed. Scoring guides and the accompanying materials are developed by an expert group using example responses collected during small-scale trialing. These are subject to ongoing review and are reviewed in the light of the field trial data.

4.6 Stimulus and Item Material: An Overview

4.6.1 Stimulus Characteristics, Selection, and Development

Stimulus materials are the essential core of many assessments, and the format, type, length, and content of stimuli can vary depending on the role they play. In the case of PIRLS, the stimuli are the reading passages that contain the text, images, and information that students read in order to respond to the items. In TIMSS, stimulus materials are the combination of text, images, and data that both contextualize items and provide information needed in order to respond to items. In ICCS, stimulus materials perform a similar role to those in TIMSS (except in the context of civic and citizenship education rather than mathematics and science). In ICILS, the stimulus materials provide both the real-world context across the narrative theme of modules and provide information that may be required to support completion of the tasks.

With the increasing use of computer-based assessment, some stimuli are now being developed exclusively for use on computer. These include all materials for ICILS, the reading passages developed for the computer-based version of PIRLS (ePIRLS), the problem solving and inquiry tasks (PSIs) in TIMSS, and the computer-enhanced modules in development for ICCS. In each case, these stimulus materials need to include interactive functionality that extends beyond what can be achieved in paper-based stimuli. While the nature of this functionality varies according to the assessment constructs being measured (as specified in the assessment frameworks) there are common criteria used to evaluate the viability of computer-based stimulus materials. They need to employ features that are accessible to students and that reflect current conventions of interface design. Furthermore the stimuli need to represent plausible (i.e., not contrived) uses of the technology in context, and operate in a broader test narrative for the students in which the use of and the reason for using the technology are apparent without the need for further explanation. Where computer-based stimuli are used, these considerations are, in addition to those that relate to the content and presentation of stimulus materials, necessary in the selection and evaluation of all stimulus materials regardless of their medium.

The development or selection of passages (texts) or other stimulus materials, such as the creation of a context within which a set of items is presented, can be the most challenging aspect of the development cycle when the relatively abstract descriptions of the assessment in the framework are operationalized. Assessment items flow from

good stimulus materials. When sourcing, selecting, and refining stimulus materials, assessment developers keep in mind the types of item that they can see will flow from the materials.

Stimulus selection and development is challenging when an assessment is being developed for use in a single jurisdiction. In the case of ILSAs, the challenge is greater as the material is scrutinized internationally. First and foremost, it must be clear which aspect of the assessable domain the stimulus is targeted at. In TIMSS, for example, the topic, and content domain (as specified by the assessment framework) can be established using the stimulus context, whereas the cognitive domain assessed will more typically be instantiated through the item. In ICCS, while some stimuli are clearly associated with a given content domain, there are also stimuli that are used to introduce civic and citizenship scenarios that elicit items across a range of content and cognitive domains.

In PIRLS, stimulus material is classified as having a literary or informational purpose. The requirement is for texts which are “rich” enough to withstand the sort of scrutiny that comes with reading items. “Rich” texts are those which are both well-written and also engaging for young students. In all assessment contexts, the selection of reading passages includes consideration of a broad suite of criteria. These include, for example, the degree to which the content of the material is: appropriate for the target grade level, inclusive, culturally sensitive, and unlikely to cause distress to readers. While this poses a challenge in local or national contexts, the challenge is significantly increased when selecting texts that are appropriate to use across a broad range of cultures. While these challenges are common when selecting stimulus materials for any ILSA, they are greatest in reading assessments such as PIRLS where there is a balance between maintaining the integrity of an original self-contained text and evaluating its appropriateness for use across countries. In PIRLS, representatives of participating countries are encouraged to submit texts they feel may be suitable; country engagement at this stage helps to ensure that literary and information texts are characteristic of the material read by children of this age around the world. Other texts are sourced by assessment experts within or associated with the international study center. Texts are submitted in English, although this may be a translation from the original. Information texts are likely to be written specifically for this assessment, and generally draw from a range of sources. In PIRLS 2016, literary texts included contemporary narrative and folk tale. Information texts were diverse and included varied purpose, layout, and structure.

In other studies, country representatives are also invited to submit stimulus materials and even ideas for materials. However, in studies other than reading there is also greater flexibility in adapting stimulus materials to ensure they are suitable for use.

An important element of the development process is to obtain the perspective of participating countries on stimulus materials during a review stage. In PIRLS, this takes place before items are written. This is necessary in PIRLS because each passage is the stimulus for a large number of items and the viability of passages must be considered before engaging in the substantial work of creating the items for each passage. Inevitably, there is considerable diversity in the viewpoints expressed. Material that is regarded as well aligned with the needs and expectations

of one country may be seen by others to be: too challenging, uninteresting, or too unfamiliar; too familiar and commonly used in classrooms; or culturally inaccessible or inappropriate. There is a high level of attrition: material “falls” at this stage and is not developed further. This is not simply a case of identifying the most popular or least problematic material; developers need to be sure that the material to be considered further has the potential to form the basis for a robust assessment. The process of “text mapping” is a means of evaluating whether a prospective text is likely to function well at the item writing stage. In text mapping, the characteristics (e.g., length, genre, form, reading load), core and secondary content and themes (explicit and implicit where relevant) of a text are listed and described. In addition there is some explication of the content and focus of items (with reference to the assessment framework) that the text naturally suggests.

In ICILS, where the real-world context of the test modules is essential to their viability, the stimulus materials are first reviewed from this perspective. Country representatives are asked to consider, in addition to the previously described criteria, the degree to which the proposed scenarios are relevant and plausible for target grade students in their national contexts. Where the stimulus materials are shorter and with more opportunity for revision, it is feasible to review them together with their relevant items.

A feature of all IEA ILSAs is the use of expert groups, separate from and in addition to country representatives, who also review all assessment materials. The expert groups typically comprise people with specialist expertise in assessing the learning area. While the size and composition of the expert groups may vary, in most cases the members are experienced researchers with at least some experience of involvement in IEA studies (as members of national research centers, for example). Many too have experience of working on national assessments in their own countries. The expert reviews can be both electronic (i.e., where feedback is sent electronically to the international study center) and delivered in face-to-face meetings. In IEA PIRLS, where the reading texts are integral to the assessment and are large and challenging to develop, the expert group can be involved in the editing and development of stimulus materials. While it varies across studies, it is typical for the expert group to provide input into the development of the stimulus materials during the early (pre-field trial) phases of development.

4.6.2 Item Characteristics and Development

The process of item development begins once the stimulus materials have been selected and revised (although this can be an iterative process in which stimulus materials are further revised as a part of the item development process).

Items fall broadly into two categories: closed response, where the student is making some sort of selection from a given set of answers, or constructed response, where the student is producing their own response. While traditionally responses in ILSAs have been largely restricted to small amounts of text (from a word or

number through to several sentences or a worked solution to a problem), the transition to computer-based testing has brought with it an expanded set of response formats. In ICILS, for example, students create information resources such as online presentations or websites and create computer coding solutions to problems. For all constructed response items, the scoring guides are developed together with the items as the two are inextricably connected. The proportion of item types is specified in the assessment framework for each study (see Sect. 4.6.3).

4.6.3 *Item Types*

Item type is generally defined by the response the student must give. Among the most recognizable “closed” item types is multiple-choice, when the student must select the correct option (the “key”) from a set of alternatives (the “distractors”). It is generally accepted that there should be at least three distractors and these should be plausible but definitively wrong. The key should not stand out in any way from the distractors (such as by being much longer or shorter than the distractors). The development of good quality multiple-choice items is harder than it may first appear, especially in ensuring that all the distractors are plausible. What makes a distractor plausible in the mind of a student is usually that student has a particular misconception, either in comprehension of the stimulus or related to their understanding of the assessment domain that leads them to the incorrect response. The nature of these misconceptions may vary according to the nature and contents of the stimulus material and the nature of the domain. For example, in PIRLS, distractors may represent an incorrect reading of the text, the imposition of assumptions based on students’ typical life experience that are not represented in the text, or the retrieval of inappropriate information from a text. In TIMSS, ICCS, and ICILS, distractors may more commonly represent misconceptions relating to the learning area. Some of these may reflect misconceptions that are well-documented in the research literature and others may represent misconceptions or process errors that are plausible based on the content of an item and the knowledge, understanding, and skills required to reach the solution. Considerable effort is undertaken to create plausible distractors. This involves test developers responding to the item from the perspective of the students, including considering the types of misconceptions that a student may have when responding. However, it is also possible to create distractors for which the distinction between the distractor and the correct response is too subtle to be discernible by students. In these cases, even though the distractor is irrefutably incorrect from an expert perspective, the capacity to discern this is beyond the reach of the students, and consequently many high achieving students believe it to be a correct answer. The empirical analysis of multiple-choice questions following the field trial allows for a review of the degree to which the distractors have been more plausible for lower achieving students and less plausible for higher achieving students.

Other closed item types, where the student is not developing their own response but indicating a selection in some way, include sequencing, where a series of statements describing or referring to a sequence are put in order, and other forms of “sorting” or matching where students indicate which pieces of information can be best matched together. Computer-based test delivery in PIRLS, TIMSS, and ICCS allows for the use of a greater range of closed item formats than can be easily developed on paper. In particular, many forms of “drag and drop” items can allow students to respond by sequencing, sorting, counting and manipulating elements. ICILS also includes a suite of closed format computer skills tasks in which students are required to execute actions in simulated software applications. Such tasks are closed, in that the response format is fixed and restricted, but from the perspective of the student the item functions as if students were working in a native “open” software environment (see Fraillon et al. 2019 for a full description of these tasks).

Constructed response items require the student to generate the content of their response. These can vary in length from a single character (such as a number) through to words, equations, and sentences. ICILS includes authoring tasks that require students to “modify and create information products using authentic computer software applications” (Fraillon et al. 2019, p. 49). These tasks can be, for example, the creation of an electronic presentation, or an electronic poster or webpage.

The scoring guide for a constructed response item is, from a content development perspective, an integral component of the item itself. As such, the scoring guide is developed alongside a constructed response item and then refined as evidence is collected. The process that the item is addressing is identified in the scoring guide, along with a statement of the criteria for the award of the point(s). The guide typically includes both a conceptual description of the essential characteristics of responses worthy of different scores as well as examples of student responses that demonstrate these characteristics. There is ongoing refinement, often following piloting, and examples are included of actual student responses. Scoring guides are incorporated into the iterative item review process when there may be further refinement.

In ICILS, the scoring guides for the authoring tasks comprise multiple (typically between 5 and 10) analytic criteria. These criteria address distinct characteristics of the students’ products and each has two or three discrete score categories. In most cases, these criteria assess either an aspect of the students’ use of the available software features in the environment to support the communicative effect of the product or the quality of the students’ use of information within their product. Despite the differences between the analytic criteria used in assessing the ICILS authoring tasks and the scoring guides developed for shorter constructed response items, the approach to the development of both types is the same. It relies on consideration of the construct being assessed as described by the assessment framework and the applicability of the guide to be interpreted and used consistently across scorers. In the case of ILSA, this includes across countries, languages, and cultures.

4.7 Phases in the Assessment Development Process

There is no single common set of activities in the development of ILSA assessments. In each study, what is completed and how is determined by the characteristics, scale, and resources of the study. In spite of this, there is a set of phases that are common to the development of assessment materials in all ILSAs. During each phase, review activities are conducted by national experts, expert groups, and the content developers, applying the quality criteria we described in Sect. 4.5.

4.7.1 Phase 1: Drafting and Sourcing Preliminary Content

This first phase of development is characterized by creative problem solving and breadth of thinking. For studies such as PIRLS or ICILS, where the texts or contexts must be confirmed before detailed item development can begin, this first phase focuses on sourcing and locating texts and conceptualizing and evaluating potential assessment contexts. In studies such as TIMSS and ICCS, in which the stimulus materials can be developed in smaller units (or testlets) with their related items, it is possible to both source and develop stimulus and items in this early phase.

During this phase, contributions from country representatives are actively encouraged and expected. Where face-to-face meetings occur it is common practice to include some form of assessment development workshop with country representatives followed by a period in which submission of texts, stimulus and assessment materials, and ideas are invited. For computer-based assessments, country representatives are typically presented with a demonstration of the testing interface and examples of items that the interface can deliver. They are then invited to propose and submit ideas and storyboards for computer-based items rather than fully developed assessment materials. Any project expert group meetings during this phase will include evaluation and development of content.

4.7.2 Phase 2: Item Development

The item development phase begins when any necessary texts, contexts, and stimuli have been selected. In this phase, the emphasis is on developing the item content and scoring guides that, together with the stimulus material and contexts comprise the draft assessment instrument. In the case of computer-based assessments, this may also include development of new item formats that support assessment that has previously not been possible. For example, in ICILS, the concept for a fully-functional visual coding system as part of the assessment was first proposed and developed in this early phase of item development.

This phase typically includes the opportunity for extensive contribution to and review of the materials by country representatives and the external experts involved in stimulus and text development. The item development phase can include any or all of the following procedures.

- Item development workshops with country representatives can be conducted early in the item development phase. The process of working in small but international teams means that different perspectives are shared and assumptions may be challenged. Whilst the working language is English, at this stage concerns about translation may emerge. Discussion of the issue across countries may ensure a resolution is identified but, in some cases, it will be found that the issue is irreconcilable and the material does not progress further in development. At this point, all items are in a relatively unpolished state, the focus being on identifying the potential for items to be developed rather than on creating the finished product.
- Piloting (and/or cognitive laboratories), in which draft assessment materials are presented to convenience samples of test-takers from the target population (usually students from countries participating in the given study) for the purpose of collecting information on the students' experiences of completing the materials. The nature of piloting can vary across projects. In some cases piloting is conducted using cognitive laboratory procedures during which students complete a subset of materials and then discuss the materials (either individually or in groups) with an administrator. In other cases test booklets are created for larger groups of students to complete in a pilot with a view to undertaking some simple quantitative analyses of item performance or to collect constructed responses from which the scoring guide can be refined and training materials developed. Piloting is of particular value when developing materials using new or changed constructs or to evaluate the test-taker experience of new item formats. This latter use has become particularly relevant in recent years, as many ILSAs transition from paper-based to computer-based formats. Where possible, piloting should be conducted across languages other than the language in which the source materials are developed (US English for all IEA studies) and can provide some early information about issues that may exist in translation of materials and the degree to which materials under translation have maintained their original meaning.
- Desktop review by country representatives and external experts is often conducted as part of the item development process. Where piloting provides information on how test-takers respond to the assessment materials, the desktop review complements this by providing expert feedback on the technical quality of the material (such as the quality of expression, clarity, and coherence and accuracy of the material), the targeting and appropriateness (such as cultural appropriateness) of the material across a broad range of countries, and how well the material represents the content of the assessment framework and the equivalent areas of learning across countries. Typically this review involves providing country representatives and experts with access to the materials (either as electronic files or through a web-based item viewing application) and inviting critical review of the materials (items, stimulus, scoring guides, and contexts). While it is possible to invite an

open review of the materials (in which respondents complete open text responses) it is common practice to structure the review so that respondents provide both some form of evaluative rating of each item and, if appropriate, a comment and recommendations for revision.

- Face-to-face meetings to review materials with country representatives and other experts are essential in the quality assurance process. While these can take place at any time during the item development cycle, they most frequently occur near the end of the process as materials are being finalized for the field trial. At these meetings, all assessment materials are reviewed and discussed, and changes are suggested. Where possible it is common to have external experts review materials in sufficient time before a face-to-face meeting with country representatives to allow for the materials to be refined before a “final” review by country representatives. A feature of IEA studies is the value placed on the input of country representatives to the assessment content. One manifestation of this is that IEA studies routinely include a meeting of national research coordinators as the final face-to-face review of assessment materials before they are approved for use in the field trial.
- Scorer training meetings occur before each of the field trial and main survey. Feedback from the scoring training meetings, in particular for the field trial, lead to refinements of the scoring guides. In most cases, the scoring guides for a study are finalized after the scoring training meeting, taking into account the feedback from the meetings.

4.7.3 Phase 3: The Field Trial and Post Field Trial Review

The field trial fulfils two main purposes:

- to trial the operational capability within participating countries
- to collect evidence of item functioning of newly developed materials.

The field trial is held approximately a year before the main survey. The size of the field trial sample will vary across studies depending on the number of items and test design in each study; however, it is usual to plan for a field trial in which no fewer than 250 students complete each item within each country (or language of testing within each country if feasible). The processes undertaken in preparation for the field trial, during the administration and afterwards in the scoring and data collection phases, mirror what is to be done during the main survey. As well as obtaining item level data, evidence collected from students may include their preferences or whether or not they enjoyed specific parts of the assessment. In PIRLS, for example, students use the universally recognized emoji of a smiley face to indicate their level of enjoyment of the passages. Item functioning is calculated, based on classical test theory and item response theory (IRT) scaling.

Following the field trial and the review of the findings by country representatives and expert groups, the final selection of material to be included in the survey is made. At this stage there are a number of considerations. Test materials must:

- meet the specification described in the assessment framework in terms of numbers of points, tests, item types, and so on;
- maximize the amount of information the test provides about the students' achievement across countries by presenting a large enough range of difficulty in the test items to match the range of student achievement;
- provide a range of item demand across sub-domains or passages, with some more accessible items at the start;
- discriminate adequately (i.e., the performance of high- and low-achieving students should be discernibly different on each item);
- contain new material that is complementary with material from previous surveys that has been brought forward, ensuring adequate and optimum representation of the construct and a balance of content (for example, a balance in male and female protagonists in reading passages);
- using evidence derived from the field trial, show sufficient student engagement and preferences;
- based on evidence from the field trial, show adequate measurement invariance for each item across countries (measured as item-by-country interaction); and
- demonstrate scoring reliability in the field trial.

The field trial also provides the first opportunity in the instrument development process for the scoring guides to be reviewed in the light of a large number of authentic student responses across countries and languages. This review allows the assessment development team to:

- check and when necessary refine the descriptions of student achievement included in the scoring guides in the light of actual student responses;
- refine the scoring guides to accommodate any previously unanticipated valid responses; and
- supplement the scoring guides with example student responses that are indicative of the different substantive categories described in the guides and develop scorer training materials.

In the lead-up to the main survey in each IEA study, national research coordinators meet to assess the data from the field trial and recommendations for the content of the assessment to be used in the main survey. Implementation of the decisions at this meeting can be considered to be the final step in the process of instrument development for that cycle of the assessment.

4.7.4 Post Main Survey Test Curriculum Mapping Analysis

The connection between explicit test and curriculum content within each country and its impact on the suitability of the assessment instrument for reporting national data is of particular importance in the areas of mathematics and science. This is because curricula in these learning areas often are based on sequences of learning content in the development of knowledge, skills, and understanding that are closely associated with specific topic contents. As such, results for an ILSA of mathematics and science achievement may be particularly sensitive to relative differences between curriculum topics in a given country and those in the ILSA instrument used across countries. In IEA TIMSS, this challenge is addressed after the main survey data collection using a test curriculum mapping analysis (TCMA). The TCMA is used to compare a country's performance on items assessing the skills and knowledge that experts in that country's curriculum consider are represented in their intended curriculum with performance on all items included in the assessment. A large discrepancy between the two sets of data (for example, a country having a much higher percentage correct on the items that were represented in their curriculum compared with the percentage correct on all items) would suggest that the selection of items for inclusion in the assessment was affecting the relative performance of countries, something which would be clearly undesirable. The results of the TIMSS 2015 curriculum matching exercise contributed to the validity of the assessment, indicating that there was little difference between each country's relative performance, whether performance across all items was considered or just those considered linked to the country's intended curriculum. Unsurprisingly, most countries perform a little better on items that are considered appropriate to that country.

4.8 Measuring Change Over Time and Releasing Materials for Public Information

While the process of developing ILSA assessment material is centered on a given assessment cycle, it is also conducted with consideration of what has come before and what is planned for the future.

The assessment development plan in an ILSA must take into account any secure material that was used in a previous cycle that will be reused in the current cycle, and also take into account what material from the current cycle may be held secure for future use. It is students' relative performance when responding to this "trend" material across cycles that is the basis for the reporting of changes in student performance over time. As such, it is essential that the content of the trend material fully represents the construct being measured and reported. How this is achieved varies according to the overarching instrument design. For example, in both PIRLS and ICILS, the test items are inextricably linked to their respective texts and modules. For this reason, in these studies, the items linked to a text or module typically span a broad range

of difficulty and cover a large proportion of the constructs. In effect, each PIRLS reading text with its items and each ICILS test module is designed to be as close as is possible to a self-contained representation of the whole assessment instrument. This allows for the selection of trend materials to be made by text or module. However, in PIRLS, where students read for literacy experience and to acquire and use information (Mullis and Martin 2019), it is necessary for both literary and informational texts to be included in the trend materials. In ICILS, a single module may be regarded as a proxy for the full assessment instrument, although it is typical to select more than one module to establish trends. In TIMSS and ICCS, where the items are developed in much smaller testlets, a large number of testlets are selected as trend materials to represent the construct. In all studies, the proportion of trend to new material is high to support robust measurement of changes in student performance across cycles.

An important aspect of communicating with stakeholders is the provision of released material from the assessments. This serves to illustrate how the assessment of the defined domain is operationalized and is indicative of the material seen by students; a useful and practical element, given the diversity of participating countries and the variety of assessment styles. Even though there is not the same measurement imperative for the released materials to represent the construct as there is for trend materials, material selected for release will ideally provide stakeholders with an accurate sense of the nature, level of challenge, and breadth of content covered in the assessment. In many cases, the released material is also made available for other research purposes.

4.9 Conclusions

The process of instrument development in any large-scale assessment is challenging and requires careful planning and expert instrument developers. However, development for ILSA introduces additional challenges. Ultimately the developers' aim is to produce instruments that function effectively in their role as a means of collecting data and assessing performance within many different countries, which present a variety of languages and cultures. While the developers cannot prescribe all possible uses of the assessment outcomes, they can and do ensure the quality of the instruments.

In this chapter we have described the constituent components and processes in the development of ILSA assessment content with a focus on four key IEA studies. While these four studies span six different learning areas that are approached and represented in different ways across countries, the fundamental characteristics and principles of assessment development are common across the studies. With the aim of maximizing the validity of assessments, the development process comprises phases of conceptualization, development, and finalization that are informed by the application of expert judgement and empirical data to interrogate the materials according to a range of quality criteria.

What lies at the core of the pursuit of excellence in this process is the feedback from experts who provide a broad and diverse set of linguistic and cultural perspectives on

the materials. Without these perspectives in the creation of ILSA materials it would not be possible to present materials that can be used confidently and defensibly assert that they can be used to measure student attainment within and across all of the countries that take part in each study.

As the range of countries participating in ILSA continues to increase and the transition to computer-based delivery continues, the ways in which computer-based assessment may improve and expand ILSA will continue to evolve. Computer-based delivery offers the opportunity to include a much broader range of item and stimulus formats than have previously been used on paper, with the opportunity to enhance assessment in existing domains and to broaden the range of domains in which ILSA can be conducted. However, this expanded repertoire of assessment content brings with it additional demands when evaluating the validity of assessments. In addition, the possibility of including process data to better understand and measure achievement is a burgeoning area that requires careful planning and integration into the assessment development process. As ILSA instruments evolve, so too must the evaluation criteria applied in the assessment development process to ensure that the principles underpinning the development of high quality assessment materials continue to be implemented appropriately.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA Publications.
- Fraillon, J., Ainley, J., Schulz, W., Duckworth, D., & Friedman, T. (2019). *International computer and information literacy study 2018 assessment framework*. Cham, Switzerland: Springer. <https://www.iea.nl/publications/assessment-framework/icils-2018-assessment-framework>.
- Fraillon, J., Ainley, J., Schulz, W., Friedman, T., & Duckworth, D. (2020). *Preparing for life in a digital world: International computer and information literacy study 2018 international report*. Cham, Switzerland: Springer. <https://www.iea.nl/publications/study-reports/preparing-life-digital-world>.
- IEA. (2020). Studies [webpage]. Amsterdam, the Netherlands: IEA. <https://www.iea.nl/studies>.
- Johansone, I. (2017). Survey operations procedures in PIRLS 2016. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 6.1–6.26). Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center. <https://www.iea.nl/publications/study-reports/international-reports-iea-studies/methods-and-procedures-pirls-2016>.
- Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>.
- Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2016). *TIMSS 2015 international results in mathematics*. Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center. <https://www.iea.nl/publications/timss-2015-international-results-mathematics>.
- Mullis, I. V. S., & Martin, M. O. (2019). PIRLS 2021 reading assessment framework. In I. V. S. Mullis & M. O. Martin (Eds.), *PIRLS 2021 assessment frameworks* (pp. 5–25). Chestnut Hill, MA: Boston College, TIMSS & PIRLS International Study Center. <http://pirls2021.org/frameworks/home/reading-assessment-framework/overview/>.

- Oliveri, M. E., Rutkowski, D., & Rutkowski, L. (2018). *Bridging validity and evaluation to match international largescale assessment claims and country aims*. Research Report No. RR-18-27. Princeton, NJ: Educational Testing Service. <https://doi.org/10.1002/ets2.12214>.
- Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences. *Large-scale Assessments in Education*, 2, 5. <https://doi.org/10.1186/s40536-014-0005-4>.
- Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25, 54–67. <https://doi.org/10.1006/ceps.1999.1020>.
- Saeed, S., & Zyngier, D. (2012). How motivation influences student engagement: A qualitative case study. *Journal of Education and Learning*, 1, 2. <http://www.ccsenet.org/journal/index.php/jel/article/view/19538>.
- Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research*, 51(2), 161–179. <https://doi.org/10.1080/00131880902891305>.

Liz Twist is joint head of the Centre for Assessment at the National Foundation for Educational Research (NFER) in England. She has been involved with IEA's Progress in International Reading Literacy Study (PIRLS) since 2000 and has led a range of large scale national assessment development projects.

Julian Fraillon is the Director of the Assessment and Reporting (Mathematics and Science) Research Program at the Australian Council for Educational Research (ACER). Julian is the Study Director of the IEA International Computer and Information Literacy Study (ICILS) 2018 and was the Study Director of the inaugural ICILS in 2013. He has been the Assessment Coordinator of the IEA International Civic and Citizenship Education Studies (2009, 2016, and 2022) and directs ACER's work on the Australian National Assessment Program studies of Civics and Citizenship and ICT Literacy. He is a member of the IEA Progress in International Reading Literacy Study (PIRLS) Reading Development Group.

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.