Localisation in International Large-scale Assessments of Competencies: Challenges and Solutions

Britta Upsing¹, Gabriele Gissler¹, Frank Goldhammer¹, Heiko Rölke¹, Andrea Ferrari²

¹German Institute for International Educational Research, Schloßstraße 29, 60486 Frankfurt am Main
www.tba.dipf.de
²cApStAn Linguistic Quality Control, Chaussée de la Hulpe 268, 1170 Bruxelles, www.capstan.be
upsing@dipf.de, gissler@dipf.de, goldhammer@dipf.de, roelke@dipf.de, andrea.ferrari@capstan.be

Abstract

International comparative studies like the Programme for International Student Assessment (PISA) pose special challenges to the localisation of the test content. To allow for comparison between countries, the assessments have to be comparable with respect to measurement properties. Therefore, internationalisation and localisation are crucial steps to guarantee test equivalence across countries. The localisation of test items is different from the localisation of web-based contents or software as the test content has to be authentic within a country while the test’s measurement properties have to be comparable across countries. Using the PIAAC study (Programme for the Assessment of Adult Competencies) as an example, this paper describes all stages of the localisation process for an international large-scale assessment. The process ranges from the development of source items to translation, adaptation of layout issues and meta-data adaptations. The paper concludes with a discussion of lessons learned and open questions.

1. Localisation in large-scale assessments

Most software or website localisation projects have the "ultimate aim of releasing a product that looks like it has been developed in country" (LISA 2003, p.11). This aim is reasonable for many instances of localisation. However, when moving to international large-scale assessment studies (studies that aim to compare skills or competence levels for given populations across countries, with a view to e.g. informing education policies), localisation is subjected to the primacy of comparability of assessment results, which may conflict with the aim of making a localised product look like it was developed in the target country itself. Unlike other localisation projects, localising assessments has to be undertaken with an eye on the comparability of multiple target versions of assessment instruments (e.g. tests). If translated tests behave differently in different countries (e.g. the difficulty varies across language versions), the significance of the research is at stake. This article will describe this potential conflict between authenticity and comparability when localising large-scale assessments on the basis of a case study.

In the remaining part of section 1, we will define large-scale assessments and add the most important details regarding the case study; this is followed by an overview of the particularities of localisation in large-scale assessment compared to web or software localisation processes. In Section 2, we will describe how these challenges can be met and show practical examples from our case. Section 3 will give an overview of the lessons learned.

1.1 What is large-scale assessment?

Policy makers around the globe need internationally comparable information about the outcomes of their education systems, information on what pupils know, and an overview of the skills and competencies of their adult workforce. This need has led to the introduction of international large-scale assessment studies, and since their implementation, localising the test content has become an important issue in the field.

In the current context, the term large-scale assessment (LSA) refers to national or international assessments that serve to describe population characteristics with respect to educational conditions.
and learning outcomes, e.g. the competence level in a particular population. Basically, LSA studies are used for monitoring the achievement level in a particular population, for comparing assessed (sub)populations, and also for instructional programme evaluation. Such assessments may form the basis for developing and/or revising educational policies.

The International Association for the Evaluation of Educational Achievement (IEA) was one of the first organisations to implement international LSA studies to assess student achievement across countries. In 1995, IEA implemented TIMSS (Trends in International Mathematics and Science Study) to assess student achievement in mathematics, just to mention one example (Mullis et al. 2009). The most widely known LSA study is the Programme for International Student Assessment (PISA) by the Organisation for Economic Co-operation and Development (OECD). The first PISA cycle took place in 2000; cycles are repeated every three years. By 2012, more than 70 countries will have participated in PISA. PISA intends to measure the knowledge and skills of fifteen-year-old students and thus make inferences on the performance of the participating countries' education systems (OECD 2010). A very first step in the shift to computer-based assessment was made in 2006 when three countries took part in the computer-based assessment of science. In 2009, participating countries had the option to evaluate the digital reading skills of their students, and a more substantial shift to the computer-based test mode was taken. 19 countries opted for this assessment (OECD 2011).

There have also been several attempts to measure the competencies of adult populations (cf. Thorn 2009): In 1994, the OECD introduced the first cycle of the International Adult Literacy Survey (IALS) to obtain information about adult literacy (prose literacy, document literacy, and quantitative literacy) in participating countries and two more rounds followed (1996 and 1998). Altogether 22 countries participated in this survey. The OECD Adult Literacy and Lifeskills Survey (ALL) builds on the results of this study and provides an international comparison of literacy, numeracy and problem-solving skills in 12 countries. It took place between 2002 and 2006. This study is now followed by the Programme for the International Assessment of Adult Competencies (PIAAC), an international large-scale survey that assesses the skills of a representative sample of adults in 25 countries.

This paper will use the example of PIAAC to describe the localisation process in LSA studies. Like PISA, PIAAC is an OECD study. PIAAC is supposed to help governments to receive "high-quality comparative information regarding the fundamental skills of the adult population" (Schleicher 2008, p. 628). The target population consists of 16-65 year old adults. The project is run by an international consortium (that includes the authors of this paper) that is responsible for enabling the local project teams to conduct the study in their respective countries. The implementation of PIAAC started in 2007. The field study1 took place in 2010; the main study will be carried out in 2011 and 2012. Results will be published in 2013. PIAAC-tests are subdivided into three different subject domains: literacy, numeracy and problem-solving in a technology-rich environment. In each of the domains, the competencies of the test participants are assessed by a number of test items2 of varying difficulty.

There is also a need to localise instructional programmes. Therefore, such programmes need to be translated into a local language. Additionally, the materials need to be adapted to local conditions. This process is called localisation. In the case of PIAAC, the localisation process includes the adaptation of the items to the local culture and context. The localisation process is also important for ensuring the validity and reliability of the assessment results. By ensuring that the assessment items are culturally and contextually appropriate, the results can be more accurately interpreted and compared across countries.
questionnaire which collects background information about the test participant. The sample includes 5000 completed interviews per country. PIAAC is a household study: the interview and the test itself take place in a respondent’s home (Thorn 2009). PIAAC is the first international LSA study that is completely computer-based, and therefore the first study to meet the specific challenges resulting from this test mode. Other studies are likely to follow this trend (e.g. PISA 2015).

As previously mentioned, localisation is an important issue because all assessment instruments (i.e., tests and questionnaires) have to be made available in the national language(s) of every participating country. PIAAC and other LSAs are challenged by localising the test items while maintaining the comparability of assessment results across countries and languages. This will be further elaborated in the next section.

1.2 Particularities of internationalisation and localisation in assessments

The localisation of LSA boils down to two questions: What exactly does it mean to internationalise and localise a test? How is this different from other localisation projects?

Adaptation of test items can occur in two scenarios and is not limited to large-scale assessments. In the first scenario, a test is originally developed for a specific language and its specific national context. Using the test internationally is not an issue when developing the test items. If, later on, the need arises to adapt the test for a new culture and language, the goal may be to obtain strict comparability, or the source test may just serve as the blueprint of a new test. This means that test developers have to decide "whether test adaptation is the best strategy" (Hambleton 2002, p. 65). In the second scenario, which is typical in the LSA context, the intended use of the test in an international comparison is a crucial factor right from the outset of developing the test. This is to ensure "that a person of the same ability will have the same probability of answering any assessment item successfully independent of his or her linguistic or cultural background" (Thorn 2009, p.8). Hence, in this second scenario, internationalisation plays an important role in making sure that the adaptation of the test will be feasible.

For computer-based tests, linguistic, cultural and technical aspects have to be taken into account to create "internationalised" source items. The following definition by Schäler (2007, p.40) is applicable for the internationalisation of LSA studies as well:

"Internationalisation is the process of designing (or modifying) digital content (in its widest sense) . . . to isolate the linguistically and culturally dependent parts of an application and of developing a system that allows linguistic and cultural adaptation supporting users working in different languages and cultures."

From a conceptual point of view, this means that source item content has to be created that is meaningful and authentic in all target cultures, as well as easily translatable. From a technical point of view, software developers have to make sure that translators can easily edit all adaptable content.

In a second step, the adaptable content has to be localised. Localisation is defined by Schäler (2007, p.40) as follows:

"Localisation is the linguistic and cultural adaptation of a digital product or service to the requirements of a foreign market and the management of multilinguality across the global, digital information flow."

In the context of LSA, not all of these factors play an important role. While Schäler emphasizes the adaptation for the target culture and making sure that the product works in the target culture, in the context of LSA, it is important that test items remain comparable across different language versions. The creation of test items for an international comparative test is thus highly demanding. On the one hand, it is important that the items are authentic within a country; on the other hand, they have to be comparable across countries. This is one of the crucial aspects that differ from other localisation processes, resulting in a multi-step adaptation process.

A second aspect deals with the material that has to be localised. In a computer-based test like PIAAC,

---

1 It should be noted though that there is a paper-based component for test participants that are not familiar with using a computer.

2 "Source" and "target" are used in this paper in the usual meaning in the translation context: the source text (or in our case the source item) refers to all aspects of an item, i.e., text, graphic elements, scoring information etc., which are being translated and/or adapted. The target text (or target item) is the translated and adapted version of the source text (source item).
localisation is not limited to the content of a test item. Meta-data like material related to the correct and incorrect responses of test items will have to be adapted as well. This is an aspect that plays a key role in the localisation process of computer-based LSA. In computer-based tests this meta-data will have to be changed in the system itself to enable automatic scoring (detailed information on this process follows in section 2.2.2).

Section 2 will explain how these two aspects are tackled in the LSA study PIAAC.

2. Case study: Localising PIAAC assessment instruments

Section 1.2 showed that the context of LSA places special requirements on the localisation process. In PIAAC, this challenge was met by first internationalising and then localising the test content. Section 2.1 describes how this was done by first creating 'internationalised' source versions of test items, while section 2.2 contextualises the insights into the localisation process itself with a focus on quality assurance.

2.1 Internationalising test items

Before the item development process can start, the "competence" that shall be measured by these items has to be defined. Basically, a competence is a theoretical construct that is used to explain and predict individual differences in behavior. Most educational LSA studies target the assessment of individual differences in competencies like "reading literacy" (in broad terms: how well can the test participant read and understand text?) or "numeracy" (again in broad terms: how well can the test participant deal with mathematical demands?). Defining the construct is a complicated process and "construct equivalence in the languages and cultures of interest" has to be kept in mind (Hambleton 2002, p. 65). Once the construct is specified and refined by an international expert group, the experts derive an "assessment framework" on the basis of the construct definition (cf. Kirsch 2001). This assessment framework explains how the test and task characteristics are related to the construct definition, and it provides systematic information about the required combinations of task characteristics to cover the construct. The creation of items can start once the assessment framework is set. In all LSA studies mentioned in chapter 1.1, the source items (see Figure 1 for an example) are created in English. They form the basis for the later localisation process. Throughout the entire item development process, the international perspective takes an all-pervasive role and several qualitative control mechanisms are in place to make sure that linguistic and intercultural aspects are considered from as many linguistic and cultural perspectives as possible. A detailed description of how such a process can be established can be found in McQueen and Mendelovits (2003). When the source items are developed, the focus is already on authenticity and comparability. The processes involved in ensuring that authentic and comparable items are created will be explained in section 2.1.1 and 2.1.2.

2.1.1 Authenticity of item content

In most software or web localisation projects, authenticity is the "ultimate aim" (LISA 2003, p. 11) as the localised projects are supposed to look like they were developed in the target country itself. For LSA studies, this means that test items should be authentic. These items should represent demands that are common and typical within a country. Furthermore, items should include task requirements that are encountered by members of the target population in their daily life. Real-life scenarios, however, are different across countries: a Japanese scenario may not be authentic in Chile. For instance, an item that asks the test participant to do a Google search and to evaluate the search results may be very authentic in many countries, but it is unfamiliar to most Koreans (where the Google search engine is hardly used). The second goal in LSA studies, i.e. comparability between localised versions, might be compromised if an item's context is familiar to some countries' populations but completely unknown in others. All localised versions of an item should function like the source version of the item, thereby yielding a high level of psychometric comparability across localised versions. The major goal is that an item has the same degree of difficulty for all countries and measures the respective construct equally well across all countries.

Hence, when item developers create the source version of a test item, they try to look for the lowest common denominator. This holds the risk of creating item material that is "bland" because the common denominator is too low. As a compromise, the following approach as used for the PISA reading assessment may be reasonable:

"The aim (...) was not to produce an instrument whose content and contexts were completely familiar to all participating
students, but, as far as possible, to control the occurrence of unfamiliarity so that no single cultural or linguistic group would be placed at a disadvantage." (McQueen and Mendelovits 2003, p. 216)

Item developers thus need to be careful when their items refer to national aspects, e.g. certain locations, institutions, education systems, currencies etc., as this raises many questions: Is the aspect known in all participating countries? Does the level of familiarity have an impact on the difficulty of the task? Is this aspect a fundamental for covering the construct?

For example, items that include aspects concerning a particular national education system raise problems even if every country might be able to localise the provided information. Educational terms (e.g. community college) can have different meanings in different countries - and be completely unknown in others. Another issue that could make a test item less authentic in some countries is any reference to the climate or weather in relation to different seasons/months. Though a scenario involving a summer party taking place in July is realistic in Europe, this scenario is not plausible in Australia.

Decisions on how to ensure authenticity have to be made on a case-by-case basis and alternative solutions are possible. Item developers could decide to replace the national reference with a fictitious name, and consequently standardise the required level of the tested persons' ability to abstract (e.g. in PISA, zed is the fictional currency unit). If the source version is not standardised in this way, item developers have to indicate to translators how to deal with this issue (e.g. if standardisation is recommended, translators might be advised to "find an equivalent institution in your country" or if standardisation is not recommended, they might be asked to "use the existing name of the institution although this institution is unknown in your country"). In most LSA studies, item developers are supported by international content experts and the participating countries themselves in making these decisions and in selecting or designing suitable items (cf. McQueen and Mendelovits 2003).

In PIAAC, similar measures were taken to control the degree of unfamiliarity across countries. Domain expert groups were set up to represent a wide range of languages and cultures. These expert groups were responsible for creating the assessment framework, which served as a basis for creating items. The item developers created items that simulate authentic real-life scenarios. The experts checked these items keeping an eye on familiarity across cultures. The selected items were presented to representatives of the participating countries, who were given the opportunity to check early versions of the items for cultural bias. Only those items that were accepted by countries were translated and used for the field test. Following the field test, items that worked inconsistently across countries were dropped or modified before being included in the main study.

### 2.1.2 Further measures for enabling comparability

To avoid item translations that could jeopardise comparability between localised versions, several measures related to linguistic and layout issues can be implemented when preparing the source items:

1) Careful linguistic construction of the source text to ensure translatability

2) Guidelines informing translators about the degree to which they can adapt translations to their countries

3) Central control of the layout of the item

4) Control of adaptable parts of an item

To ensure translatability, item developers refer to a number of general guidelines. For example, they should only use idiomatic speech in the source version of an item with great care, as it could be very difficult to find adequate formulations for each of the target languages. Also, it might be difficult to find adequate translations of things like proverbs. Item questions should not be directed at the "level of nuances of language" (McQueen and Mendelovits 2003, p.215). Generally, the passive voice should be avoided because it does not exist in all languages (Hambleton 2002, p. 71).

Item creation must be accompanied by detailed translation guidelines for preparing the subsequent localisation process, otherwise comparability between target versions would be questionable from the outset because translators for different languages might assume different degrees of "translating freedom". These guidelines should answer all questions that a translator may have regarding the adaptation of specific item content ("Can I adapt the number format to the number format that is used in my country?", "Can I adapt the name of the institution?"
...). In addition, guidelines should
provide general instructions for the translation of assessment items. This can include explaining which style of speech needs to be used in certain settings, general information about translating assessment items. For example: make sure that answer choices are kept about the same lengths in the translation so that they do not become a clue to the correct answer, information about the target audience, etc. (Hambleton, Merenda & Spielberger 2005).

In PIAAC, translators received "translation guidelines" with general instructions on how to translate assessment items. A second document, the "translation and adaptation guidelines", describes the structure and content of each item as well as the correct and incorrect answers. It gives advice for translating item-specific content, e.g. on how the translator should deal with names (adapt or not?). In addition to the general translation and adaptation guidelines, a so-called verification follow up form (VFF) is used to organise and control the localisation process. The VFF is a spreadsheet containing all text elements of an item and related instructions, including precise translation/adaptation advice relating to specific text elements (what should be adapted, what should not, how to understand ambiguous or difficult terms, pointers on consistency both within and across units, etc.). The VFF serves as a means to document all comments and successive translated versions of each item as it goes through the different phases of the localisation process: double translation and reconciliation, verification, country's post verification review, layout adaption, finalisation (for more details, see section 2.2.1).

The context of LSA studies may involve specific requirements regarding item layout when designing the source versions. Item developers want to be in control of the item layout across language versions as the position of information that is crucial for completing a task may affect item difficulty (Freedle 1997). This is the case when scrolling is required to see all of the text included in an item (for example in a stimulus that imitates a webpage); or when a long text is divided into several columns. To ensure comparability in these cases, it may be important that the starting position of text elements like headlines, paragraphs or the location of the correct response is exactly the same for all language versions. This could be solved by designing the source version in a way that precludes the introduction of cross-country variability in critical properties of the text layout. Therefore, the item editing software should allow for defining the absolute position of each element on the screen. In PIAAC, the CBA ItemBuilder was used as a tool for developing the source version of test items. The concept of the CBA ItemBuilder is to enable item writers to design and edit computer-based test items with the aid of a graphical editor that can easily be used by non-IT-specialists. The different components of an item can be positioned in the drawing area. The item writer has full control over the absolute size and position of the different components because each element can be aligned pixel by pixel on the screen. Consequently, the location of these elements cannot be changed when the text is translated. In anticipation of layout problems that could occur after translating the English source version to the different languages, the size of each text field was not only made as large as necessary for the English text, but was enlarged by approximately 1/3 to have enough space for languages that require more space for the same content, e.g. German or Russian.

Finally, with regard to the subsequent localisation process, it needs to be decided which components of the source items need to be adaptable, and which should be static across language versions. Basically, only those elements which are meant to be translated or adapted during the localisation process should be adaptable. Otherwise, comparability may be compromised due to uncontrolled changes.

An item usually consists of graphical and textual elements. For computer-based items, these textual elements can also include meta-data like scoring information. All textual elements need to be adaptable for translating the content to the target language. In addition, one could also think of adapting the graphical elements of a test item. For example, this would be necessary when adapting an item that simulates a website to a right-to-left written language system. To achieve an authentic context for this language version, not only does the text need to be adapted, but also the text layout and the website structure.

In PIAAC, none of the participating countries used a right-to-left written language system; therefore only textual elements were made adaptable. Also, all countries were supposed to use the same images as the source item. As a consequence, textual and graphical elements needed to be technically separable. Moreover, graphics should not contain any textual elements but if needed were superimposed by textual elements. Even symbols were to be avoided or at least checked in terms of their international
suitability.

The software that was used for building the source versions allowed the separation of the entire textual content from the graphical representation of an item, and to export this adaptable content as an XLIFF file. Later on in the localisation process, this XLIFF file was used for translation purposes. Once the text had been translated and validated, the XLIFF file was reimported to the test item.

The finalised internationalisation process results in a set of carefully checked and reviewed source items. These items serve as a basis for the localisation process, which will be described in section 2.2.

2.2 Localising test items
The localisation process consists of several steps to obtain items that can function comparably across countries as well as being authentic within a country. The content - mostly text - included in the item has to be adapted, but in several cases the layout or the scoring has to be adapted as well.

Section 2.2.1 will describe the adaptation and quality assurance procedures involved in adapting the textual content, section 2.2.2 will describe the layout adaptations, and section 2.2.3 will explain why metadata such as the scoring of an item may have to be adapted as well, and how this can be done.

2.2.1 Localising the content
The International Test Commission Test Adaptation Guidelines (cf. Hambleton and de Jong 2003, p. 129) ask for a highly sophisticated translation procedure:

"D.5 Test developers/publishers should implement systematic judgmental evidence, both linguistic and psychological, to improve the accuracy of the adaptation process and compile evidence on the equivalence of all language versions."

The translations should correctly deliver the content, be authentic and fluent, and at the same time they must not change the psychometric properties of the item. Thus, for LSA it is recommended to set up rigorous translation procedures that involve more than one translator for the adaptation of test items. Also, one individual can hardly meet the required translator's profile:

"There is considerable evidence suggesting that test translators need to be (1) familiar with both source and target languages and the cultures, (2) generally familiar with the construct being assessed, and (3) familiar with the principles of good test development practices." (Hambleton 2002, p. 62)

For LSA, the double-translation design is recommended. Double-translation means that two translators create two independent translations of the source text. This is followed by a "reconciliation", which consists of merging the two independent translations into one target version. As Grisay (2003, p. 228) puts it:

"equivalence of the source and target languages is obtained by using three different people (two translators and one reconciler) who all work on the (sic!) both source and the target versions."

In general, the idea is bringing together linguistic, domain and assessment experts that work as a team in creating the best possible target version.

In PIAAC, double-translation and reconciliation were carried out by the project teams within countries and the translation efforts were subsequently checked by a "verification" process provided by the international consortium in charge of the project. Specially recruited and trained verifiers checked both formal correspondence of target version to the source version and fluency/correctness in the target version, striving to achieve an optimal balance between these two goals, which sometimes pull different ways (e.g. maintaining the order of presentation of the information within a sentence or passage - versus opting for a more "natural" order in the target language). They also check whether the above-mentioned layout and adaption guidelines are followed. Verification was followed by a discussion with the reconciliation team. An optical layout check was also necessary because the translation often had an impact on the layout. This was then followed by testing of the scoring mechanism (cf. chapter 2.2.3) and finally by testing the integrated assessment tests.

For this multi-step localisation process, extensive documentation of all changes and comments is indispensable, as also highlighted by the

XLIFF is the abbreviation for XML Localisation Interchange File Format. It is a standard file format which permits making adaptable data editable and manageable within a localisation process (Savourel et al. 2008).
International Test Commission Test Adaptation Guidelines\(^6\) (Hambleton and de Jong 2003, p. 130):

"I.1 When a test is adapted for use in another population, documentation of the changes should be provided, along with evidence of the equivalence."

In PIAAC, so-called Verification Follow Up Forms (VFF\(^7\)) were used, which contained the aforementioned translation and adaptation guidelines and provided space for discussion for the different people involved in the translation process. The verifier who checked the reconciled version could add comments and recommendations to one or several parts of the translation, and the country's reconciliation team could respond by accepting or refusing the verifier's recommendations. In the VFFs, the different players could also explain the reasons and motives for their decisions. Thus, for each country detailed documentation was generated that contained a summary of the decisions made for every single localisation issue. Errors or changes that were valid for all countries were compiled in a special "errata sheet" available for all countries.

In practice, translators (or reconcilers or verifiers) were only able to translate the text derived from the test items and made available for them in the aforementioned XLIFF file. Everybody involved in the translation process could preview the English source version of the test item on a web-based Item Management Portal. More importantly, they were also able to interact with the item in the way the test participant would during the test (e.g. they could answer the item, click on links within the stimulus, see all different pages that were included in items that simulated webpages etc.). After translating an XLIFF file or after correcting a translated version, it was possible to upload the translation to the portal and preview the translation there. For the translation of the XLIFF text, the Open Language Tool (OLT)\(^8\) was used. The OLT includes a Translation Memory, which helps to maintain consistency across test units.

2.2.2 Localising the layout

Layout adaptations became necessary after translating despite all efforts made during the internationalisation process described in section 2.1.

Country teams were required to check all their items for potentially corrupt layout and report these issues to the consortium, which then tried to adapt the layout as required by the country. This resulted in a protracted exchange of communication between all partners involved until all problems were taken care of.

As mentioned in section 2.1.2, the source version provided extra space to accommodate languages whose translations take up more space than English does. In several cases, the allocated space was still insufficient and had to be extended (or resulted in a smaller font). For languages that took up less space than the source version, the layout had to be adapted in a few cases as well.

In a few isolated cases, graphics had to be exchanged in a localised item for authenticity reasons (for example, an image that shows bottles had to be exchanged when the beverage itself was not known in the country or carried specific connotations).

Also, justified text - which looked like, for example, an authentic newspaper article in the source item - looked unusual in some translations because the languages had much longer word lengths than the English original. This problem was solved by hyphenating words. In such cases, hyphenated text was not included in text that was crucial for answering the item. All of these issues (and more) were discussed and checked by item experts to ensure that they would not compromise cross-country comparability.

2.2.3 Localising the scoring

In a computer-based test, a respondent can provide answers in several ways: response types can include multiple choice, short text entry, numeric entry, selection of radio buttons or combo boxes, text field entry, highlighting text, marking graphical objects or cells, and many more. The entries given by the respondent then have to be scored. Scoring items means that a score is assigned to the test participant's response. The score is defined by a scoring rule, which relates (ranges of) responses to scores. Automatic, machine-based scoring requires defining scoring rules within the system. Manual scoring, by human experts, relies on scoring guidelines including scoring rules and assignments of typical responses to scores.

---

\(^6\) These guidelines were set up to support test item developers when adapting test instrum ents (Hambleton and de Jong 2003).

\(^7\) cA pStA n, a linguistic quality control company, was responsible for the generation of VFFs and for the general translation and adaptation procedures. Their verifiers were responsible for checking the translated versions produced by the country teams.

\(^8\) The OLT is an open-source tool that is available online (The Source for Java Technology Collaboration n.d.).
Most response types, with the exception of free text entry, can be automatically scored by a computer system in a straightforward manner. Automatic scoring can be more efficient than human scoring as the time-consuming work by human scorers is not necessary. Whenever adaptive testing is used, automatic scoring becomes a pre-requisite.

In a test that has to be translated, adaptation of the scoring usually does not pose any difficulties for response types such as multiple-choice or marking graphical objects or cells. Here it is most important that the text is translated. The choice of correct or incorrect responses usually does not change their location (in most languages) and no further adaptation is necessary. However, some scoring information is language or country specific and has to be prepared in a way that allows for its localisation.

One example is the scoring of numeric responses, for instance in the case of items involving currencies. An item might ask the respondent to calculate the price of a purchase, e.g. "This radio costs 30 dollars. How much does it cost when a 10% discount is given?". The correct response is "27 dollars" in our example. If the price of the radio and the correct response are not adapted in a country with a different currency (for example, Japan where 1 USD = approx. 80 Yen), the item context is no longer authentic. In PIAAC (in contrast to PISA where the fictional currency zeds is used, as mentioned earlier), real currencies were retained, with guidelines for adaptation. In such a case, the localisation of the scoring content becomes inevitable and the defined correct response will have to be changed in the system.

Localising scoring rules of numeric entry items requires not only the definition of the correct number(s) but also decisions about acceptable spelling formats for numbers (e.g. with respect to the kind of decimal separator). Although there are international standards for number formats defining the spelling of numbers by country, it may be too strict to accept only responses as correct if they adhere to these standards. Given considerable variability in the usage of number formats within countries (and even within test participants), a more lenient scoring approach that accepts alternative number formats was judged to be more appropriate for PIAAC.

In PIAAC, complexities also arose from the adaptation and localisation of the highlight response mode. For the highlight response mode, the respondent has to mark the correct answer in the stimulus text to indicate his or her answer. Here is an example to illustrate this and to explain how the scoring mechanism is designed in PIAAC:

![Figure 2: Preview of a sample highlight item](image)

*Adaptive testing means that the number pattern of correct and incorrect responses of a respondent has an impact on the difficulty of the next test items that are presented. The idea is that a test taker that repeatedly shows low level skills is more likely to receive easy items, while a respondent that shows high level skills is more likely to receive difficult items. So in Computerized Adaptive Testing the item difficulty is tailored to the individual's performance level. Too hard and too easy items which would not contribute to a reliable measure are avoided. In an adaptive test, the upcoming item or set of items is selected adaptively based on the performance shown in previous items. In some instances, e.g. for selecting the first item (set), additional contextual information (e.g. educational level) may be used as well (Wainer 2000).*
The respondent is given a text and he is asked to highlight information in the text to give his answer. The question refers to the stimulus text and asks: "What is the latest time that children should arrive at preschool?" (cf. Figure 2).

Text that is not included in any of the text blocks CAN be selected. It is a part of the so-called maximum correct response (which is "Please have your child here by 9:00 am.").

After the translation of the text content, it is important that the text blocks are redefined as well, because they are language-dependent and thus unlikely to match the source version in terms of size and location. In PIAAC, this followed reconciliation and subsequent check by the verifiers. For the adaptation of the text blocks the "Translation Textblock Editor" was used, a standalone tool derived from the CBA ItemBuilder mentioned above.

Countries could not define new text blocks or delete text blocks, but they were able to adapt the content of the text blocks according to their needs. This process required several informed decisions about how to localise the scoring rule in a comparable way as illustrated with a simple example.

**Question: What does the text say about how much computer scientists earn?**

**Stimulus text:**

"Computer scientists under 30 typically make more than the average salary for their age from day one."

In the source version of the item, the minimum correct response text block consists of "more", "than" and "average". Once the text is translated into German, "average salary" becomes "Durchschnittsgehalt". Should test participants receive a correct response when they only highlight "Durchschnitts" (which represents "average" in this
Scoring experts within the countries had to find answers to many scoring-specific questions, e.g. how to deal with compound words; how to deal with endings (e.g. should inflections be included in the minimum correct response?); is the correct response still comparable to the source version when the target version includes significantly more words in its minimum correct response?

After the field test, the text blocks could be re-adapted if the field test results showed that items in one country behaved differently from items in other countries. The localisation of scoring was a difficult task for the countries.

3. Lessons learned and open questions

This paper, so far, has given a brief introduction to LSA projects and discussed the role of localisation in the area of LSA studies. As previously described, localising tests for international LSA studies poses specific challenges that are not necessarily encountered in other localisation processes. One of the main differences concerns the struggle between authenticity and comparability when localising, and the adaptation of scoring information. By describing a real scenario, we examined how these aspects are dealt with in practice. PIAAC is special in its own right because it is the first international LSA study that is completely computer-based (with a paper-based option for inexperienced computer users). The multi-step procedure that was implemented to manage these difficulties poses some open challenges for future studies. Many of these challenges result from the shift to a computer-based test mode and can be classified into two categories: firstly, new difficulties concerning the localised content and, secondly, and more importantly, difficulties regarding the internationalisation and localisation process when trying to master both complexity and quality assurance. These challenges will be described in the following paragraphs.

With regard to the test content, special linguistic difficulties arise within the new field of test items that simulate technology-rich environments (web pages, software tools,...). The question of authenticity arises when web content is translated into languages with a low population of speakers, like Estonian: a stimulus mimicking a web page might be considered as inauthentic if completely translated into languages for which only limited content is available on the web. Also, people in some countries typically do not use their national language as an application interface language (for example because the localised interface was only introduced very late and people were already used to working with an English interface). Hence, the question arose as to whether the interface language should be translated or not. Similar concerns can arise for languages with different fonts (for which it is difficult to translate URLs in a web browser). Not translating this content might make the item more difficult for respondents who are less familiar with using a computer (or do not speak English). Translation, however, might make the item inauthentic, which might have an impact on the difficulty of the item as the technical terminology might be less familiar to the test taker. Similar problems can arise when tests are translated into minority languages (like Valencian or Basque). Even though inauthenticity might be less of a problem for speakers of these languages (as many of them are familiar with using their language in new contexts), there might also be an impact on the difficulty of the tasks. These issues and their influence on an item's validity of measurement will have to be discussed further in the future.

The shift from a paper-based to a computer-based test mode has a significant impact on the adaptation processes. One big difference compared to the adaptation process for paper-based tests is the separation of adaptable content from static, non-adaptable content. On the one hand, this makes the process more complicated and requires many case-by-case decisions. On the other hand, it automatically brings many issues to light that would not necessarily be (knowingly) identified during a localisation process for paper-based tests (Should the inline formatting be exactly the same across languages? Can the font size be changed? What degree of freedom is allowed for changing layout?). In addition, the computer-based mode of test items technically facilitates the direct comparisons of localised test items. Hence, the shift presents a challenge as well as an opportunity for making localisation issues more visible than before.

This also leads to the problem of finding the right balance between flexibility and control. In PIAAC, a conscious decision was made not to allow the countries or the software to make any changes to the layout. As previously mentioned, this was helpful because the consortium (and the item developers) maintained control over the location of the text. On the other hand, it is questionable whether it would not...
have been preferable to allow for more decentralised layout adaptations. If the size of a text box automatically adapted to the length of the translated text, many of the manual adaptations of the items (bearing the risk of introducing new errors) could have been avoided. Especially for languages like Korean and Japanese, it would also be helpful if countries were granted more flexibility to adapt some selected elements of the layout manually. Line space, for example, had to be doubled for Korean because the Korean characters become illegible with the default line space set in the source items. For Japanese, line breaks were also an issue: there are no blanks between characters and text is usually justified. When designing the translation process and the software tools for the translation process, these requirements should play a role from the very beginning and be a part of the items' internationalisation process. Certain countries would thus gain access only to selected layout elements that could not be dealt with during internationalisation.

The adaptation process for computer-based tests also requires the ability to integrate two additional steps into the localisation process, i.e. layout and scoring adaptations. Defining the sequence of the adaptation steps becomes a challenging task in such a complex process. For example, allowing any linguistic changes to be made after the completion of scoring and layout adaptations means that these adaptations have to be re-checked. An ideal approach would be to first complete all linguistic changes, and secondly resolve all layout issues. The scoring should be adapted at the very end. Since the localisation of automatic scoring rules is a new area in LSA, and the consequences of scoring adaptations are not visible in the item itself, countries need to test scoring carefully following a test plan.

It also became clear that it is important that all people involved in the adaptation process are able to interact with the item in the same way as the test participant. This also became apparent for the scoring mechanism, for the adaptation of which it was crucial to be able to test all changes by trying to give correct and incorrect responses. Countries received detailed test cases from the consortium giving the correct or incorrect responses for the source version, which could then be adapted by the country and checked on the Item Management Portal by giving the required response. The portal then gave feedback on whether the response was correct or incorrect. This allowed for immediate feedback on whether the adaptations (of e.g. text blocks) resulted in the desirable scoring behaviour. This procedure - testing while adapting - made the scoring adaptation process efficient for countries because they received immediate feedback for any scoring adaptation decision.

Another challenge regarding the efficiency of the localisation process refers to the question of who should make adaptations, i.e. whether certain adaptation steps should be centralised and done by experts in the consortium, or de-centralised and become the responsibility of the national teams. For instance, at the beginning of the project, the consortium tried to give countries the freedom of adapting their numeric scoring. This decision was made because the people in the national teams would be able to decide if items that include currencies should be adapted or not (cf. previous section). However, it soon became clear that it was not efficient to teach this complicated adaptation procedure to all countries: input was needed from numeracy experts to decide whether changing a currency number would change the item's psychometric properties, such as difficulty, as well as from technical experts to implement the changed scoring rule. In PIAAC, this process was then modified and centrally organised: the consortium and the numeracy expert group made recommendations and gave feedback regarding certain problematic items, the country groups made sure that the items were authentic for their country, and the consortium made the technical implementation. A conclusion from the PIAAC case study is that it is more efficient to implement technically difficult adaptations centrally, after countries have provided input as regards authenticity.

One step to allow for more de-centralisation and transparency regarding content decisions would be to give countries more, and broader, information as a basis for making decisions and finding solutions during the localisation process. For instance, as a future enhancement of the PIAAC approach, one might try to make information regarding localisation issues available and useable across countries, so that each country team can gain a new cross-country perspective and is able to compare different localisation problems and solutions. The bundling of information could result in a more consistent approach and increased quality. Many localisation problems do not only exist for one language but across languages. In these cases it would be very helpful for a country's translation and localisation team if they had an overview of all the localisation problems that emerged for an item in other countries.
Furthermore, they could check whether they might have a similar problem that they are not yet aware of. In addition, once a problem is identified, they could directly check solutions other countries had found for a similar problem and use these solutions as a guideline for their own decision. A technical solution for such a centrally available cross-country information and documentation pool would be needed for the localisation process.

Source version management is a difficult issue in an adaptation process that includes many different partners in many different countries. Even though the source items, after "internationalisation", are supposed to be final prior to starting the localisation process, several issues are only found once countries have started on their translations, and more are found through the verification procedure. One problem regarding the file-based solution in PIAAC was that every time a new version became available, countries had to download this version and check that this was the latest version. A lot of these issues can probably be avoided by advance translation: This is done in PISA, for example, where two source versions are created: a French source version is developed in parallel with the English source version. At least some of the issues that concern the translatability of items can thus be identified in advance. There are fewer errors when the source versions are released for translation by the countries (Grisay 2003). Still, it is likely that not all problems can be found, even by using advance translation. Source version management itself could be technically supported by using a content management system, which would prevent subsequent errors caused by miscommunication between partners or overlooking changed material.

The question of source version management leads to the question of translation version management. The multi-step localisation procedure also made it difficult for countries to translate because they had to consult and edit a lot of material for translation. This should be reduced so that cycling between many documents is not necessary anymore; a technical solution should be found. A first step in this direction has been made with the framework of PISA 2012 computer-based testing, whereby item-specific translation/adaptation guidelines and comments by the different players (translators, reconciler, verifier, country post-verification reviewer) are carried within the XLIFF file rather than being presented in a separate monitoring form.

4. Conclusion

As described in this article, many problems have to be tackled in LSA studies that are not usually present in localisation processes where comparability does not play a role. In particular, localisation in LSA studies deals with balancing between authenticity in each country and comparability across countries. To handle this challenge, a multi-stage translation and verification approach is pursued, including:

- Preparing internationalised test material
- Localising content (text, images)
- Localising layout
- Localising meta-data, e.g. scoring rules.

Still, several aspects can be transferred to other localisation processes as well. For instance, the issue of version management is of general importance, as well as the question of when to test a localised version. Other domains for which the quality of translations is highly critical might also benefit from the multi-stage translation and verification process that is used for LSA studies. Similarly, the question as to which adaptations should be done, and by whom, is also relevant in all localisation processes.

On the other hand, LSA studies can take more advantage of the advances made by the localisation industry. As LSA studies are shifting from paper-based assessment to computer-based assessment, the time seems right to move towards commonly used standards and tools. In PIAAC, the first steps in this direction have been taken by introducing the XLIFF standard as a basis for the translation and by requiring countries to use a translation memory (TM) aware translation tool such as the OLT. Nevertheless, not all of the new possibilities have been tried yet. Another promising approach is to put more emphasis on source content quality assurance.

References


