frontiers | Frontiers in Education

ORIGINAL RESEARCH
published: 11 May 2022
doi: 10.3389/feduc.2022.885825

Linguistic, Contextual, and Experiential Equivalence Issues in the Adaptation of a Performance-Based Assessment of Generic Skills in Higher Education

Jani Ursin[1], Heidi Hyytinen[2], Kaisa Silvennoinen[3] and Auli Toom[2]*

[1] Finnish Institute for Educational Research, University of Jyväskylä, Jyväskylä, Finland, [2] Centre for University Teaching and Learning, University of Helsinki, Helsinki, Finland, [3] Department of Education, University of Jyväskylä, Jyväskylä, Finland

This qualitative study investigated the various linguistic, contextual, and experiential equivalence issues embedded in a performance-based instrument aimed at assessing generic skills in higher education. A rigorous translation and adaptation process (American English to Finnish) was conducted on one instrument, namely *Collegiate Learning Assessment* (*CLA+*) *International*. The data were obtained from cognitive laboratories (*n* = 13), with think-alouds and follow-up interviews conducted among Finnish undergraduate students. Content logs were created, and the data were analyzed thematically. The findings revealed that linguistic and contextual equivalence issues were more prominent than experiential ones. The findings underline how important – and potentially problematic – it is for a test to measure the same construct in a different language and culture. To achieve adequate measurement equivalence, a detailed qualitative analysis of linguistic, contextual, and experiential equivalence should be conducted as part of test adaptation.

Keywords: cross-cultural adaptation, equivalence, higher education, translation, performance-based assessment, generic skills and competences

OPEN ACCESS

Edited by:
Maria J. Hernandez-Serrano, University of Salamanca, Spain

Reviewed by:
Jacqueline P. Leighton, University of Alberta, Canada
Edith Braun, Justus-Liebig-Universität Gießen, Germany

****Correspondence:***
Jani Ursin
jani.p.ursin@jyu.fi

Specialty section:
This article was submitted to Higher Education, a section of the journal Frontiers in Education

Received: *28 February 2022*
Accepted: *25 April 2022*
Published: *11 May 2022*

Citation:
Ursin J, Hyytinen H, Silvennoinen K and Toom A (2022) Linguistic, Contextual, and Experiential Equivalence Issues in the Adaptation of a Performance-Based Assessment of Generic Skills in Higher Education. Front. Educ. 7:885825. doi: 10.3389/feduc.2022.885825

INTRODUCTION

International (comparative) assessments of learning outcomes such as generic skills have become popular in many countries. Several such assessments have been conducted by the Organization for Economic Co-operation and Development (OECD), including the Programme for International Student Assessment (PISA), and the Programme for the International Assessment of Adult Competencies (PIAAC). In higher education there have also been initiatives to measure learning outcomes from a comparative perspective, notably the OECD Assessment of Higher Education Learning Outcomes (AHELO), which investigated what students at the end of their first (bachelor level) degree know and are able to do (see Tremblay et al., 2012). The AHELO was a performance-based assessment that included two complementary components to assess generic skills: selected-response questions (SRQs), and an open-ended performance task (PT). Nonetheless, the AHELO measurement of generic skills has been criticized as being inadequately contextualized and as disproportionately "American" in an international context, with consequent issues of content validity and reliability (Tremblay et al., 2012; Shavelson et al., 2019). In fact, such challenges

are typical of cases in which assessments are developed in a given country and then transferred to other contexts.

Despite this, adapting an existing test has many advantages as opposed to developing and validating a completely new instrument, bearing in mind the resources required, such as high-level expertise in the skills or knowledge being assessed, deep contextual and cultural understanding, and time and money (Ercikan and Lyons-Thomas, 2013; Schendel and Tolmie, 2017). Adapted tests are therefore used frequently, especially in cross-national comparative studies (Hambleton and Patsula, 1998; Hambleton and Lee, 2013). Nevertheless, it is impossible to evaluate such studies and draw conclusions on the findings without a carefully implemented and fully reported translation and adaptation process, with careful attention to equivalence issues (van Widenfelt et al., 2005; Borsa et al., 2012). To successfully adapt a test instrument from one cultural setting to another requires more than merely translating the original test into the target language on a word-for-word basis (Borsa et al., 2012; Ercikan and Lyons-Thomas, 2013; Ercikan and Por, 2020; Ronderos et al., 2021). Typically, the translation and adaptation process includes phases of translation, reconciliation, back translation, expert reviews, pretesting, and evaluation of the final structure (e.g., Karlgren et al., 2020). Especially in selected-response question formats, the evaluation of the final structure is often conducted *via* factor analysis. However, such a quantitative approach is insufficient as a sole indicator of validity when applied to inherently complex performance-based test instruments that include detailed instructions and a number of complex reference documents. Qualitative analyses are therefore needed in efforts to adapt performance-based assessments of generic skills, having in view possible culturally embedded meanings that are difficult to detect by purely statistical means (Ronderos et al., 2021). In fact, most previous studies on the adaptation and validity of performance-based assessments of generic skills in higher education have been quantitative in nature (e.g., Zlatkin-Troitschanskaia et al., 2019; Kleemola et al., 2021), with only a few qualitative analyses of validity (Schendel and Tolmie, 2017; Karlgren et al., 2020).

The present study sought to fill this gap. It did so by analyzing students' response processes while they were carrying out assessment tasks, relating these to the various linguistic, contextual, and experiential equivalence issues embedded in a performance-based instrument aimed at measuring generic skills, namely the *Collegiate Learning Assessment* (*CLA+*) *International.* This measures undergraduate students' generic skills, including problem solving, reasoning, critical reading and evaluation, and written communication. The instrument was originally developed in the United States, then implemented in the Finnish higher education context. Here, linguistic equivalence refers to the notion that words should mean the same thing in the target language. Contextual equivalence refers to an instrument and its parts having the same relevance and being understood in a similar fashion irrespective of the context, and experiential equivalence means that instrument and its parts need to have a similar intention or function in the target culture. Our research question was as follows: *What kinds of linguistic, contextual, and experiential differences can be found in the adaptation of the CLA+ International into Finnish higher education?*

THE CROSS-CULTURAL ADAPTATION OF A TEST INSTRUMENT

In order to obtain reliable results and improve the validity of generic skills measurements, the instrument must be adapted with great care to ensure its usability in a new cultural context. It has often been noted that adapting a test instrument from one cultural setting to another requires more than just translating the original test into the target language on a word-for-word basis (Hambleton, 2005; Borsa et al., 2012; Ercikan and Lyons-Thomas, 2013). Hence, test adaptation aims "to maintain equivalence in content and cultural meaning between the original and the translated/adapted test, thus fostering the comparability of scores across individuals from [. . .] different cultural groups" (Hernández et al., 2020, p. 390).

Test translation and test adaptation are intertwined as concepts and processes. However, according to Hambleton (2005), they refer to different things. *Test adaptation* has been understood as a broad term referring to the various activities that are needed when preparing to use a test in another language and culture; by contrast, *test translation* can be seen as merely one phase of this process (Hambleton, 2005; Ercikan and Lyons-Thomas, 2013; International Test Commission, 2018). This aspect is dealt with in greater detail by Ronderos et al. (2021; see also Hambleton and Patsula, 1998), with translation being seen as the creation of linguistically equivalent versions of a test, in contrast with adaptation, which involves cultural considerations such as equivalence of the construct, similarity of test administration, speed of response, and familiarity with the item format. For its part, the term *cross-cultural adaptation* can be used to describe a process in which not just language, but also other aspects related to cultural adaptation, are taken into consideration in translating and adapting a test instrument to a new cultural context (Beaton et al., 2000; Ercikan and Por, 2020).

The research literature presents a large number of guidelines and suggestions for adapting test instruments for use in another culture and for evaluating the quality of this process (see Beaton et al., 2000, 2007; Hambleton, 2005; Gudmundsson, 2009; Borsa et al., 2012; Ercikan and Lyons-Thomas, 2013; Hambleton and Lee, 2013; International Test Commission, 2018; Oliveri and Lawless, 2018; Hernández et al., 2020). Nonetheless, researchers have noted that there is no clear agreement on the ideal adaptation method (Borsa et al., 2012; Epstein et al., 2015). The test adaptation process may vary depending on the instrument and its intended use (Gudmundsson, 2009; Borsa et al., 2012; Hernández et al., 2020).

Typically, test adaptation includes the phases of (1) considering whether the measured construct can be captured by a test in another cultural context, (2) translating/adapting the test (by competent translators) and deciding on the kinds of accommodations needed in order to use the test in another language and culture, (3) evaluating the quality and equivalence of the translations, and (4) pretesting the adapted test (see

Beaton et al., 2000, 2007; Hambleton, 2005; Borsa et al., 2012). The first step refers to how far the intended construct has a similar meaning in different cultures (Hambleton, 2005). Such construct equivalence between source and target cultures is crucial, as without it, cross-cultural comparisons are impossible (Hambleton, 2005; Ercikan and Lyons-Thomas, 2013).

Secondly, test adaptation guidelines emphasize the importance of an accurate translation process. The recommendation is to use multiple trained translators who are familiar with both the source and the target languages and cultures (Hambleton, 2005; Beaton et al., 2007). To obtain translation accuracy, translators should be provided with sufficient information on the nature of the instrument being adapted (Hambleton, 2005; Arffman, 2013). In addition to forward translation, there have been recommendations also to use back translation (i.e., having the adapted test translated back into its original language) as a step to evaluate the quality and validity of translations (e.g., Beaton et al., 2000, 2007; Borsa et al., 2012). However, in their review regarding cross-cultural adaptation methods and guidelines, Epstein et al. (2015) noted that back translation has generated considerable controversy: some practitioners have regarded it as an essential part of cross-cultural adaptation whereas others make no such recommendation, especially in cases where the adaptation team speak both the source and the target language.

Thirdly, different translations made independently by translators should be synthesized and then evaluated by an expert group (Beaton et al., 2000, 2007; Borsa et al., 2012). In this way, possible equivalence issues and sources of translation/adaptation errors can be identified. The equivalence between two language versions of the test may be lacking for a variety of reasons. For example, translations may change the content or meaning of test items. In order to maintain item equivalence, it is essential to consider to what extent and in what way this change has taken place (Ercikan and Lyons-Thomas, 2013). The aim of translations is not just to find words but also expressions and concepts that have both linguistically and culturally similar meanings in the target culture (Hambleton, 2005). Literal translation is unlikely to be the optimal way to proceed, as it can lead to errors in terms of test content, linguistic, or cultural factors (van Widenfelt et al., 2005; Karlgren et al., 2020) – a phenomenon also referred to as "unwanted literal translation" (Arffman, 2012). Indeed, all translation requires some degree of adaptation, as translations depend on characteristics of the target language including "its interplay with the intended meaning of a test item and the features of the source and target culture and population" (Ronderos et al., 2021, p. 66).

Finally, the adapted test and its functionality should be pretested in practice within the intended target group (Beaton et al., 2000, 2007). In addition to examining the content and characteristics of test items, pretesting makes it possible to evaluate other factors related to the test, such as the functionality of the instructions (Borsa et al., 2012; Hambleton and Lee, 2013). Indeed, pretesting an instrument is particularly crucial for performance-based assessments of generic skills, which include open-ended questions (such as performance tasks), detailed instructions, and several qualitative background documents on which students must base their answers. Properly conducted pretesting of a performance-based assessment will help to reveal possible sources of error that might threaten the validity of the instrument. In the performance-based assessment of generic skills it is pivotal that the questions should not contain unfamiliar words or complicated structures that would produce comprehension problems (Johnson et al., 2009).

As indicated above, equivalence is imperative in translation and in the adaptation of a test from one culture to another. Equivalence refers to the requirement that different language versions should be comparable to each other, and measure the same construct (Arffman, 2013). Overall the literature presents various forms and categorizations of equivalence. According to Karlgren et al. (2020; see also Borsa et al., 2012) one needs to check whether words mean the same thing (*semantic equivalence*), whether colloquialisms or idioms need to be replaced (*idiomatic equivalence*), and whether the "same" word holds a different conceptual meaning in the culture (*conceptual equivalence*). *Experiential equivalence* is also important. This means that items have to be replaced by something addressing a similar intention or function in the target culture; for example, knife and fork may need to be replaced with chopstick if that is the common utensil used for eating in target culture. Furthermore, participants in different cultures may not be equally familiar with certain test item types, such as selective-response questions (Hambleton and Patsula, 1998; Hambleton, 2005; Ercikan and Por, 2020). This aspect relates to *item equivalence* (Herdman et al., 1998). The structure of the test instrument or the way in which the test is administered are also important factors to consider from the perspective of cultural adaptation (e.g., Herdman et al., 1998; Hambleton, 2005; Schendel and Tolmie, 2017). This is known as *operational equivalence* (Herdman et al., 1998). In addition, *measurement equivalence* – meaning that the two versions should not differ significantly in their psychometric properties – is often viewed as a distinct form of equivalence (Herdman et al., 1998; Epstein et al., 2015).

Because the concept of equivalence has various forms and meanings and many of them are closely linked to each other, we see is as useful to summarize the forms of equivalence that we apply. These are: (1) *linguistic equivalence*, incorporating elements from semantic, idiomatic and conceptual equivalence and referring to the notion that words should mean the same thing in the target language, (2) *contextual equivalence*, meaning that an instrument and its parts have the same relevance and are being understood in a similar fashion irrespective of the context, and (3) *experiential equivalence*, based on the notion that an instrument and its parts should have a similar intention or function in the target culture.

COLLEGIATE LEARNING ASSESSMENT INTERNATIONAL AS A PERFORMANCE-BASED TEST INSTRUMENT

This study utilized the test instrument *Collegiate Learning Assessment* (*CLA+*) *International*. CLA+ International is a subject-independent performance-based assessment developed

by the United States-based Council for Aid to Education (CAE), which measures undergraduate students' generic skills. For our part, we understand generic skills as universal expert skills needed in studies and working life. In higher education, higher-order skills such as analytical reasoning skills and problem-solving skills are typically valued more highly than practical generic skills. Performance-based assessment aims to cover generic skills in an authentic manner by giving an opportunity for students to demonstrate their skills measured in the assessment task (Shavelson, 2010; Hyytinen et al., 2021). Performance-based assessment refers to a variety of task types, such as open-ended performance-task and document-based selected-response questions. Typically, a performance-based assessment will challenge students to use their higher-order thinking skills to create a product or to complete a process (Braun et al., 2020). Indeed, a performance assessment can be viewed as "an activity or set of activities that requires test takers [. . ..] to generate products or performances in response to a complex, most often real-world task" (Davey et al., 2015, p. 10). Thus, students actively participate in the problem-solving exercise and may even learn during the performance-based assessment (cf. Kane, 2013), rather than passively selecting answers (Palm, 2008; Hyytinen et al., 2021).

In line with this definition of performance-based assessment, CLA+ International includes three components. First of all, a student has 60 min to respond to a performance task which measures analysis and problem solving, writing effectiveness, and writing mechanics. The performance task includes an open-ended question in which students are asked to produce a justified solution to a presented real-life problem, utilizing in their written response different source materials from an online Document Library. Responding to the performance task requires students to simultaneously use a range of generic skills, as they need to analyze and evaluate information, make conclusions, and provide evidence for their own solution or recommendation (Shavelson, 2010; Zahner and Ciolfi, 2018; Hyytinen et al., 2021). In this study, the performance task was about comparing life expectancies in two cities, and students had to consider whether some measures were needed to increase the life expectancy in one of the cities. In their responses, the students had to present a solution to the problem and to give a recommendation for possible measures. The source materials that students needed in order to formulate their response contained five different source documents: a blog text, a transcribed podcast, a memorandum, a newspaper article, and infographics (see Ursin et al., 2021).

Thereafter, students had 30 min to answer 25 selected-response questions. Ten of the questions were relevant to the background document, which dealt with the secretion of proteins in the brain. These questions measured scientific and quantitative reasoning. A second set of ten questions, based on a letter about nanotechnology sent by a reader to an imaginary journal, measured critical reading and evaluation. The last five questions, which related to an opinion piece on women in combat, assessed the student's ability to analyze arguments, including possible logical fallacies. At the end of the test, the students filled in a background information survey (Ursin et al., 2021). Because the test tasks are still used internationally, the performance task and selected-response questions used in this study cannot be published or described in a more detailed manner. However, similar test tasks are presented by Shavelson (2010) and Tremblay et al. (2012).

TRANSLATION AND ADAPTATION OF COLLEGIATE LEARNING ASSESSMENT INTERNATIONAL TO THE FINNISH CONTEXT

The CLA+ test was translated into Finnish. The translation and adaptation of the test progressed through four main steps as specified in the guidelines of the International Translation Committee (International Test Commission, 2018; cf. Hambleton and Patsula, 1998). In the first phase, the test was translated from English into Finnish by a qualified translator with knowledge of large-scale assessments in the field of education. In the second phase, two trained translators in Finland reviewed, confirmed and, if necessary, proposed changes or corrections to the translations independently of each other. In the third phase, the project team in Finland decided on the final versions of the translations on the basis of the translators' proposals. The reconciled revisions were then verified by the test developer in the United States. The translated test was then pretested in Finnish in "cognitive laboratories," ensuring that the translation and adaptation phase had not changed the meaning, the level of difficulty, or the internal structure of the test (see Ursin et al., 2021). The suitability of the test for the Finnish context was ensured in detail. The translation and adaptation of the test instrument did not include a phase of back translation, since, as noted above, previous studies (e.g., Epstein et al., 2015) have indicated that it may not be a necessary step, especially if the research personnel speak both the source and the target language, which was the case in this study.

AIMS, MATERIALS, AND METHODS

The main aim of the study was to identify various equivalence issues in adapting the CLA+ International instrument. More specifically, we focused on the differences to be found in the adaptation of CLA+ International from the United States context to Finnish higher education (cf. Hambleton, 2005; Borsa et al., 2012; Karlgren et al., 2020), in line with our categorization of issue types. We see the differences as involving:

(1) Linguistic issues (whether words mean the same thing in the target language);
(2) Contextual issues (whether an instrument or its parts has the same relevance and are being understood in a similar fashion irrespective of the context);
(3) Experiential issues (whether the instrument or its parts have a similar intention or function in the target culture).

The data came from 13 cognitive lab events with think-alouds and follow-up interviews, conducted on a target group of students (**Table 1**). The participants, who all were

TABLE 1 | Demographics of the participants in the cognitive labs.

Gender	Field of study	Type of higher education institution
Male	Humanities and Arts	University
Female	Humanities and Arts	University
Male	Science	University
Female	Science	University
Male	Humanities and Arts	University
Female	Health and Welfare	University of Applied Sciences
Female	Health and Welfare	University of Applied Sciences
Female	Health and Welfare	University of Applied Sciences
Female	Services	University of Applied Sciences
Male	Engineering, manufacturing, architecture, and construction	University of Applied Sciences
Male	Arts	University of Applied Sciences
Female	Arts	University of Applied Sciences
Female	Arts	University of Applied Sciences

white Caucasians, represented two large multidisciplinary higher education institutions in southern Finland. One of the institutions was a research-intensive university, and the other was a professionally oriented university of applied sciences. Participation in the research was voluntary, and informed consent was obtained from the participants. The cognitive labs made it possible to collect authentic data on participants' ongoing thinking processes and behaviors while they were working on a task (van Someren et al., 1994; Leighton, 2017, Leighton, 2019). The data were collected individually from all the participants by following a similar procedure. At the start of each lab, the participants were instructed and trained to think aloud as they were solving the tasks. Verbalization took place when the participant first completed the performance task, and thereafter during 25 selected-response questions in a secured online environment. To avoid bias in the data collection, a neutral form of the think-aloud protocol was used (van Someren et al., 1994; Leighton, 2017). It follows that the participants were not interrupted while they were performing the tasks. "Keep talking" was the only probe given during the lab if the participant was silent for a long time. The researchers sat in the back of the room and kept their distance from the participants while they completed the tasks. The neutral form of the think-aloud protocol ensured that the probing questions were not asked until the follow-up interview. In the second phase, after the think-aloud, a short follow-up interview was conducted. This included both targeted questions (based on the observations during the think-aloud phase) and general questions posed to all participants (covering notably the clarity of the instructions, the comprehensibility of the test, how interesting the test was, the strategy used for answering). The first- and second-named authors collected the data.

Each lab lasted around 2 h and was videotaped and recorded by a camera and a table microphone. In addition, notes were taken by the researchers. The verbalizations of each participant during the cognitive labs were transcribed verbatim. After that, content logs were created in which accurate descriptions of non-verbal actions, a summary of events, and transcriptions of the verbalizations of each participant were combined into one text file (Oranje et al., 2017; see **Table 2**). A content log provides an overview of the video data, and it can be used to locate sequences and events for further analysis. The log externalizes and visualizes participants' thinking processes and behaviors associated with the assessment constructs and progress in the task. A strength of the log is that it encompasses all the input provided by a test-taker, i.e., direct quotes, assertions, behavior, and actions that took place during the think-aloud process. The log includes information on the sequence, timing, and variety of the test-taker's response behaviors and actions. Furthermore, it combines both verbal and non-verbal response processes (Hyytinen et al., 2014; Oranje et al., 2017).

The transcripts and content logs were analyzed using a thematic approach in which, initially, similar notions were systematically coded under preliminary content categories. Subsequently, final categories were formed on the basis of a relational analysis (Braun and Clarke, 2006). Finally, the preliminary categories were further elaborated on a theoretical basis, with special attention paid to issues of contextual, linguistic, and experiential equivalence (Hambleton, 2005; Borsa et al., 2012; Karlgren et al., 2020). Furthermore, numbers of occurrence were calculated in order to reach an understanding of how typical a given category might be. The first and third author of this paper did the initial coding; this was then revised against the coding made by the second and fourth authors. Thereafter, the final categories were discussed and agreed with all the authors. Translated and anonymized excerpts from the cognitive laboratories were selected for each category for illustrative purposes.

RESULTS

All of the participants (n = 13) experienced equivalence issues while taking the test. The analysis identified several linguistic, contextual, and experiential issues (**Table 3**). The equivalence issues identified related mainly to how questions were formulated, and how materials were comprehended; also to how the instructions were presented, and how certain concepts were understood.

Linguistic Equivalence

Most of the issues related to linguistic differences between Finnish and English. By linguistic equivalence we refer to the notion that the meaning of the words and phrases should not have altered in the translation and adaptation from English into Finnish. One language-related difference concerned the phrasing of the questions. Efforts had been made to keep the equivalence between English and Finnish phrases as close as possible, but this occasionally created situations where it was difficult for a student to understand the translated question. An example from a cognitive lab reads as follows:

> *[The participant] reads the question, ponders for a moment what it says ("the following criteria, that is, these [criteria] except one of them, is that so?")* (SRQ item 3 – ID17).

TABLE 2 | An example of a content log (Hyytinen et al., 2021).

Time	Duration	Code of student: ID2
0:00:00–0:00:27	0:00:27	Logging into the test plus the privacy notice Glances through the privacy notice and asks how to move on. Asks the same thing also at the summary of the test. Glances through the summary of the test
0:00:27–0:08:14	0:07:47	Performance task
0:00:27–0:01:08	0:00:41	Reads and glances through the general instructions for the performance task
0:01:08–0:08:14	0:07:06	Moves to the actual performance task. First, quickly reads the task instruction and some of the documents. Moves directly to writing the answer, does not plan it beforehand. Browses the documents. Concentrates on the infographic. Using that as a basis, says that "the physical activity habits of the residents should be improved." Does not substantiate the answer more precisely, compare the information in the documents, or evaluate the reliability of their content aloud. Completes the answer in no more than 8 min and moves on to the SRQ items Written response: *The physical activities of Brookdale's inhabitants should be improved. The inhabitants must be told about a healthy diet. The education level must be improved*
0:08:14–0:59:52	0:51:38	SRQs
0:08:14–0:10:39	0:02:25	Moves to the SRQ section. Browses through the SRQ instructions. Asks for help on how to move on
0:10:39–0:15:32	0:04:53	Glances at the first question and items, then the document provided for the first SRQ set. Moves back to the first question and items, then identifies the relevant section from the source document. Compares the items to the document. Thinks aloud which item (A–D) would most weaken the main claim of the document. Says that "option A could be true based on the document, hence A is not the right answer." Selects option D. Moves to the second question

TABLE 3 | Linguistic, contextual, and experiential equivalence issues in the data (with number of occurrence).

	Linguistic equivalence (*n* = 29)	Contextual equivalence (*n* = 20)	Experiential equivalence (*n* = 4)
Questions (*n* = 19)	Phrasing of the questions (*n* = 19)		
Materials (*n* = 17)	Difficulties in understanding the text (excessive use of abbreviations) (*n* = 6)	Differences in understanding a reliable source of information (*n* = 6)	
		Proper interpretation of a figure (map) (*n* = 5)	
Instructions (*n* = 13)	Difficult linguistic expressions (*n* = 4)	Multitude of instructions (*n* = 5)	
		Usefulness of instructions (*n* = 4)	
Concepts (*n* = 4)			Difficulties in understanding the meaning of concepts in the Finnish context (*n* = 4)

The way the question was posed was not a typical way of presenting a question in Finnish, thus making it rather difficult to understand. Nonetheless, changing the formulation of question into a more "Finnish" formulation might also have impacted on the difficulty of the question (made it less difficult). Hence, no substantial changes were made to the formulation of this particular question.

Another language-related issue was how the instructions were given in the online test environment. This resulted in situations, for example, where students were unsure how to move forward in the test platform because they were confused about the linguistic expression and symbol represented (in the original) by "mark complete." The following excerpt from a cognitive lab exemplifies this:

> *[The participant] is wondering for a moment how to move forward from the instructions, until she clicks on "mark complete"* (SRQ – ID19).

"Mark complete" was initially translated into Finnish (literally) as "merkitse valmiiksi" which is not a typical (although a possible) way of expressing that one can now move on to the next page in the online platform. To make the instruction more understandable it was ultimately changed to the more conventional "valmis" (meaning "completed" in English), thus avoiding the pitfall of unwanted literal translation (Arffman, 2012).

The final language issue related to the background materials used in the test instrument. These were required for a student to answer the questions. Typically, Finnish does not use abbreviations as readily as English. One of the SRQ documents (regarding secretion of proteins in the brain) included an excessive use of difficult abbreviations (from the point of view of the Finnish language) making it challenging and occasionally frustrating for students to understand the text. This is reflected in the following excerpt from a cognitive lab: [*The participant] notes that she is too tired to concentrate on a text filled with abbreviations* – (SRQ item 1 – ID14). Although some other students also reported challenges related to the use of abbreviations, this example of frustration emerged as an extreme case, and ultimately no changes were made to the background document.

Contextual Equivalence

Several context-related equivalence issues became visible in the analysis. By contextual equivalence we refer to the requirement for an instrument and its parts to have the same relevance and

meaning in the target culture as it has in the culture of departure. The first contextual equivalence issue related to the background materials on which the students had to base their answers. On a few occasions the students wondered what a reliable source of information might actually be ultimately leading to a question whether reliable source of information is comprehended similarly among undergraduate students in United States and in Finland. One of the participants felt that some of the documents were ridiculous, almost to the point of neglecting the document altogether:

> *Wonders aloud about those [documents]; notes that the articles, on the basis of which the report should be made, seem a bit "silly." Questions the relevance of Document 4, does not find it a reliable/relevant source and is going to ignore it* (Performance task – ID12).

This had an impact on the quality of the answer, as students were informed in the instructions for the performance task that their answers would be judged on how thoroughly the information was covered, including mention of potential counterarguments. Hence, completely ignoring some of the documents might have resulted in poorer test scores. Another example relating to contextual equivalence concerned the interpretation of the figures in one of the background documents. Thus, in order to interpret one of the figures correctly, a student should be familiar with the intermediate compass points (such as South-East) on a map representing the United States. As the use of intermediate compass points to describe Finland as a country is not as typical as it is in the United States, this led to challenges for some students in attempting to answer a question. One of the students reacted as follows:

> *[R]eads the question and examines the figure. Is not sure about the compass points and comments that this task makes no sense if you don't recall [the compass points]* (SRQ item 10 – ID3).

The second contextual equivalence issue related to how instructions in the background documents were formulated. As compared to common practice in the United States, in Finland matters are typically presented without much guidance or orientation to the reader. Consequently, metatext tends to be used much less in the Finnish context than in the Anglo-American context (see Mauranen, 1993). In several background documents, multiple instructions were given, including lengthy guidance. This caused confusion to some of the participants. The following example from a cognitive lab illustrates the challenges due to the excessive use of orientation text:

> *[The participant] moves on to the instructions for the performance task, reads/goes through it. Notes that there is much to read in the instructions* (Performance task – ID13).

Another participant was uncertain whether all the instructions in the test were actually needed or relevant:

> *[The participant] is reading the privacy notice and asks [from the researchers carrying out the cognitive lab] if one can just accept it. Reads the summary of the test and asks for specifications about task duration. Asks if the instruction section can be skipped* (ID3).

Experiential Equivalence

There were also some experiential equivalence issues found in the data. By experiential equivalence we mean that the instrument and its parts should have a similar intention or function in the target culture. There were a few concepts such as "drinking water" and "ordinary diet" which might have been experienced differently by the Finnish participants. In Finland, drinking water is typically the same as tap water (which is high-quality, drinkable, and of similar taste across the country), but this is not the case in the United States. Although this point was not particularly crucial from the point of view of answering the question posed in the performance task, it might have led to a different understanding of the concept of "drinking water" depending whether one was an undergraduate student in the United States or in Finland. One of the participants pondered what the "ordinary Finnish diet" mentioned in Document 5 might actually mean (Performance task – ID13). "Ordinary" was initially translated as "tavallinen" in Finnish, but was ultimately changed to "perinteinen," meaning "traditional" in English. Nonetheless, the issue remained whether an ordinary/traditional diet means the same thing in the United States and in Finland.

DISCUSSION

By analyzing students' response processes during tasks, our study aimed to identify various linguistic, contextual, and experiential equivalence issues embedded in the CLA+ International translation and adaptation process from the United States to the Finnish context. In our study, linguistic and cultural equivalence issues emerged interestingly as more crucial than experiential ones (cf. Hambleton, 2005; Borsa et al., 2012; Karlgren et al., 2020). The issues of linguistic equivalence were associated with the formulation of questions, difficulties in understanding some linguistic expressions in the instructions of the test, and challenges in comprehending one of the SRQ documents due to an excessive use of abbreviations. Contextual equivalence issues were related to the interpretation of a figure (how to make sense of a map of the United States), and to the abundance and utility of the instructions posed in the test instrument, with difficulties also in understanding what a reliable source of information could consist of. There were only a few issues of experiential equivalence, linked to difficulties in comprehending the meaning of certain concepts (such as "drinking water") in the Finnish context. Nevertheless, our findings show that linguistic, contextual, and experiential factors need to be taken into account in interpreting the performance-based assessment of generic skills. All these aspects affect how students interpret the task, instructions, questions and materials used in assessments, and how they formulate their responses (Ercikan and Por, 2020). If students face unfamiliar or completely new ways of presenting a test in a situation, this may demand additional capacity from them, and thus influence their performance in the test.

The findings confirm previous notions of what is needful for translations: not merely to find words, but also expressions and

concepts that have both linguistically and culturally a similar meaning and intention in the target culture (Hambleton, 2005; Arffman, 2012; Ercikan and Por, 2020). An example of this occurred in the way a student was given instructions when answering questions. Because Finnish texts tend to include only minimal explicit metalanguage to orient the reader, as compared to Anglo-American texts (Mauranen, 1993; Kleemola et al., 2022), the multitude of instructions in some instances raised concern over whether the instructions actually embodied the same intention in Finnish cultural context. Some of these findings may seem minor, but they could have a considerable impact in the test situation and on students' test performance.

Our findings also indicated the extent to which equivalence issues can be intertwined. For example, how a word or term is translated (linguistic equivalence) might also change how it is understood in different contexts (experiential equivalence). An example in our data was the term "ordinary diet," which was confusing to Finnish participants, and led to a revision of the translation (to "traditional diet") in the final version of the test. One can then ask whether undergraduate students in United States higher education would understand the term "traditional diet" in a similar fashion to their counterparts in Finland. Another example of the intertwined nature of equivalence issues in our data was about how the Finnish undergraduate students can make sense of a map of United States when the use of intermediate compass points in Finland is not as typical as in the United States. While this is contextual equivalence issue (whether the map of the United States has the same meaning in Finland as in the United States) this is also "experiential" issue insofar as it relates to geography, and the large area of the American land mass, and the geographical variations it contains. This experience of United States as a country is something that the Finnish participants are lacking. Overall, our findings showed the extent to which the translation and adaptation of CLA+ International from American to Finnish context involved a process of carefully balancing between content, language, and experiential factors (see van Widenfelt et al., 2005; Karlgren et al., 2020).

Our findings contribute to the assessment literature by suggesting a need for greater recognition of equivalence and validity issues in translated and adapted performance-based assessments of generic skills in higher education. This is important in order to guarantee collecting high-quality research data. This is of crucial importance as opposed to pure selected-response questions, performance-based assessments typically include a complex set of background documents and instructions. As shown in our study also, these increase the likelihood of cultural, linguistic, and experiential equivalence issues in the test instrument (cf. Ercikan and Por, 2020). Consequently, in performance-based assessments it is crucial to identify equivalence issues if one is seeking to diminish their effect on participants' test-taking. Our findings importantly support the notion that to ensure that a test measures the same construct in a different language and culture, a qualitative analysis of equivalence issues is a necessary part of test adaptation, together with psychometric evidence (e.g., Ercikan and Pellegrino, 2017). However, it is important to note that without cognitive laboratory data, it would not have been possible to gather authentic data on participants' ongoing response processes while they worked on a task. A key observation of our study to the assessment literature is that qualitative analyses of cognitive laboratory data are of enormous help in revealing possible challenges in the validity and equivalence of an adapted test instrument. In the long run, such research is crucial for the development of the generic skills research field, which at present lacks robust replicable instruments (Braun et al., 2012; El Soufi and See, 2019; Tuononen et al., 2022).

LIMITATIONS OF THE STUDY AND FUTURE RESEARCH

The findings of our study can be used to improve the quality of a translated and adapted generic skills assessment instrument. However, certain limitations in the study should be taken into account. The first of these relates to the relatively small amount of data obtained, given that the data comprised 13 cognitive labs with think-alouds. This nevertheless resulted in around 26 h of recorded data, and it can be claimed that saturation was reached in terms of sufficiency of the data. In the future, cognitive labs could be carried out with a more versatile group of undergraduate students (i.e., from different disciplinary backgrounds), though one has to bear in mind that the setting up of cognitive labs followed by detailed analysis of the data (including the creation of content logs) requires considerable resources. A further limitation concerns the think-aloud method. It is possible that the participants' ability to verbalize their thoughts might have biased the think-aloud data. Note, however, that in order to minimize bias in the data collection, we followed a formalized procedure. This included instructions and explanations to participants on thinking aloud, a brief training session, and a neutral protocol that avoided probing questions. In this way, we endeavored to ensure the reliability of the verbal data (van Someren et al., 1994; Leighton, 2017). A third limitation is related to the three forms of equivalence (linguistic, contextual and experiential) that we used in our paper. The different forms of equivalence are intertwined to the extent that it is difficult to make a clear-cut analytical distinction between the different manifestations of equivalence. Furthermore, contextual and experiential equivalence is strongly related to the characteristics of the participant; if participants represent a sub-culture or belong to a particular ethnic group, they may have a different understanding of an instrument (or parts of it) from that of the majority of the population in the target culture. A fourth limitation is linked to the fact that the translated and adapted test instrument included only one performance task and one set of selected-response questions. A more reliable picture of the equivalence issues would have been obtained by including more than just one of each type of task. Hence, in future it would be important to study the equivalence issues pertaining to different kinds of performance tasks and selected-response questions, since these might enter into the performance-based assessment of generic skills in higher education. Furthermore, it would also be important to develop tasks in an international context by a knowledgeable team of experts, and to study whether

such tasks would include fewer equivalence issues than those developed in a single country. Note also that linguistic, cultural, and experiential equivalence issues appear to be closely bound up with the methodological and technical aspects of a test instrument (Hambleton, 2005).

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because they contain information that could potentially identify the participants, and thus compromise their anonymity. The datasets may also reveal information on the test that comes under the copyright of the test developer. Requests to access the datasets should be directed to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the Local Legislation and Institutional Requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

All authors listed have made a substantial, direct, and intellectual contribution to the work, and approved it for publication.

FUNDING

The present study was supported by the Finnish Ministry of Education and Culture. This study was carried out through the funding of a project entitled *Korkeakouluopiskelijoiden oppimistulosten arviointi Suomessa* (KAPPAS!).

ACKNOWLEDGMENTS

We would like to thank Donald Adamson for the helpful comments and proof-reading the manuscript.

REFERENCES

Arffman, I. (2012). Unwanted literal translation: An underdiscussed problem in international achievement studies. *Educ. Res. Int.* 2012, 1–13. doi: 10.1155/2012/503824

Arffman, I. (2013). Problems and issues in translating international educational achievement tests. *Educ. Meas. Issues Pract.* 32, 2–14. doi: 10.1111/emip.12007

Beaton, D., Bombardier, C., Guillemin, F., and Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *Spine* 25, 3186–3191. doi: 10.1097/00007632-200012150-00014

Beaton, D., Bombardier, C., Guillemin, F., and Ferraz, M. B. (2007). *Recommendations for the Cross-Cultural Adaptation of Health Status Measures.* Available online at: http://dash.iwh.on.ca/sites/dash/files/downloads/cross_cultural_adaptation_2007.pdf (accessed May 20, 2021).

Borsa, J. C., Damaìsio, B. F., and Bandeira, D. R. (2012). Cross-cultural adaptation and validation of psychological instruments: Some considerations. *Paideìia* 22, 423–432. doi: 10.1590/1982-43272253201314

Braun, E., Woodley, A., Richardson, J. T. E., and Leidner, B. (2012). Self-rated competences questionnaires from a design perspective. *Educ. Res. Rev.* 7, 1–18. doi: 10.1016/j.edurev.2011.11.005

Braun, H. I., Shavelson, R. J., Zlatkin-Troitschanskaia, O., and Borowiec, K. (2020). Performance Assessment of Critical Thinking: Conceptualization, Design, and Implementation. *Front. Educ.* 5:156. doi: 10.3389/feduc.2020.00156

Braun, V., and Clarke, V. (2006). Using thematic analysis in psychology. *Qual. Res. Psychol.* 3, 77–101. doi: 10.1191/1478088706qp063oa

Davey, T., Ferrara, S., Shavelson, R., Holland, P., Webb, N., and Wise, L. (2015). *Psychometric Considerations for the Next Generation of Performance Assessment.* Available online at: https://www.ets.org/Media/Research/pdf/psychometric_considerations_white_paper.pdf (accessed Feb. 17, 2022)

El Soufi, N., and See, B. H. (2019). Does explicit teaching of critical thinking improve critical thinking skills of English language learners in higher education? A critical review of causal evidence. *Stud. Educ. Eval.* 60, 140–162. doi: 10.1016/j.stueduc.2018.12.006

Epstein, J., Santo, R. M., and Guillemin, F. (2015). A review of guidelines for cross-cultural adaptation of questionnaires could not bring out a consensus. *J. Clin. Epidemiol.* 68, 435–441. doi: 10.1016/j.jclinepi.2014.11.021

Ercikan, K., and Lyons-Thomas, J. (2013). "Adapting tests for use in other languages and cultures," in *APA Handbook of Testing and Assessment in Psychology, Testing and Assessment in School Psychology and Education*, eds K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, et al. (Washington, DC: American Psychological Association), 545–569. doi: 10.1037/14049-026

Ercikan, K., and Pellegrino, J. W. (2017). "Validation of score meaning using examinee response processes for the next generation of assessments," in *Validation of Score Meaning for the Next Generation of Assessments: The Use of Response Processes*, eds K. Ercikan and J. W. Pellegrino (New York, NY: Routledge), doi: 10.1111/jedm.12256

Ercikan, K., and Por, H. H. (2020). "Comparability in multilingual and multicultural assessment contexts," in *Comparability of Large-Scale Educational Assessments: Issues and Recommendations*, eds A. Berman, E. Haertel, and J. Pellegrino (Washington, DC: National Academy of Education), 205–225. doi: 10.31094/2020/1

Gudmundsson, E. (2009). Guidelines for translating and adapting psychological instruments. *Nord. Psychol.* 61, 29–45. doi: 10.1027/1901-2276.61.2.29

Hambleton, R. K. (2005). "Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures," in *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, eds R. K. Hambleton, P. F. Merenda, and C. D. Spielberger (Mahwah, NJ: Lawrence Erlbaum), 3–38.

Hambleton, R. K., and Lee, M. K. (2013). "Methods for translating and adapting tests to increase cross-language validity," in *The Oxford Handbook of Child Psychological Assessment*, eds D. H. Saklofske, C. R. Reynolds, and V. L. Schwean (Oxford: Oxford University Press), 172–181. doi: 10.1159/000477727

Hambleton, R. K., and Patsula, L. (1998). Adapting tests for use in multiple languages and cultures. *Soc. Indic. Res.* 45, 153–171. doi: 10.1023/A:1006941729637

Herdman, M., Fox-Rushby, J., and Badia, X. (1998). A model of equivalence in the cultural adaptation of HRQoL instruments: the universalist approach. *Qual. Life Res.* 7, 323–335. doi: 10.1023/a:1024985930536

Hernández, A., Hidalgo, M. D., Hambleton, R. K., and Gómez-Benito, J. (2020). International Test Commission guidelines for test adaptation: A criterion checklist. *Psicothema* 32, 390–398. doi: 10.7334/psicothema2019.306

Hyytinen, H., Holma, K., Shavelson, R. J., and Lindblom-Ylänne, S. (2014). The complex relationship between students' critical thinking and epistemological beliefs in the context of problem solving. *Frontline Learn. Res.* 6:1–15. doi: 10.14786/flr.v2i4.124

Hyytinen, H., Ursin, J., Silvennoinen, K., Kleemola, K., and Toom, A. (2021). The dynamic relationship between response processes and self-regulation in critical

thinking assessments. *Stud. Educ. Eval.* 71:101090. doi: 10.1016/j.stueduc.2021.101090

International Test Commission (2018). ITC guidelines for translating and adapting tests (Second edition). *Int. J. Test.* 18, 101–134. doi: 10.1080/15305058.2017.1398166

Johnson, R., Penny, J., and Gordon, B. (2009). *Assessing Performance. Designing, Scoring, and Validating Performance Tasks*. New York, NY: Guilford Press.

Kane, M. (2013). The argument-based approach to validation. *Sch. Psychol. Rev.* 42, 448–457. doi: 10.1080/02796015.2013.12087465

Karlgren, K., Lakkala, M., Toom, A., Ilomäki, L., Lahti-Nuuttila, P., and Muukkonen, H. (2020). Assessing the learning of knowledge work competence in higher education – cross-cultural translation and adaptation of the Collaborative Knowledge Practices Questionnaire. *Res. Pap. Educ.* 35, 8–22. doi: 10.1080/02671522.2019.1677752

Kleemola, K., Hyytinen, H., and Toom, A. (2021). Exploring internal structure of a performance-based critical thinking assessment for new students in higher education. *Assess. Eval. High. Educ.* doi: 10.1080/02602938.2021.1946482

Kleemola, K., Hyytinen, H., and Toom, A. (2022, in press). The challenge of position-taking in novice higher education students' argumentative writing. *Front. Educ.* doi: 10.3389/feduc.2022.885987

Leighton, J. P. (2017). *Using Think-Aloud Interviews and Cognitive Labs in Educational Research*. Oxford: Oxford University Press.

Leighton, J. P. (2019). The risk–return trade-off: Performance assessments and cognitive validation of inferences. *Br. J. Educ. Psychol.* 89, 441–455. doi: 10.1111/bjep.12271

Mauranen, A. (1993). Contrastive ESP rhetoric: metatext in Finnish-English economics texts. *Engl. Specif. Purp.* 12, 3–22. doi: 10.1016/0889-4906(93)90024-I

Oliveri, M. E., and Lawless, R. (2018). *The Validity of Inferences from Locally Developed Assessments Administered Globally (Research Report No. RR-18-35)*. Princeton, NJ: Educational Testing Service, doi: 10.1002/ets2.12221

Oranje, A., Gorin, J., Jia, Y., and Kerr, D. (2017). "Collecting, analyzing, and interpreting response time, eye-tracking and log data," in *Validation of Score Meaning Using Examinee Response Processes for the Next Generation of Assessments*, eds K. Ercikan and J. W. Pellegrino (New York, NY: Routledge), 34–44.

Palm, T. (2008). Performance assessment and authentic assessment: a conceptual analysis of the literature. *Pract. Assess. Res. Eval.* 13, 1–11. doi: 10.7275/0qpc-ws45

Ronderos, N., Shavelson, R. J., Holtsch, D., Zlatkin-Troitschanskaia, O., and Solano-Flores, G. (2021). International performance assessment of critical thinking: framework for translation and adaptation. *J. Supranat. Policies Educ.* 13, 62–87. doi: 10.15366/jospoe2021.13.003

Schendel, R., and Tolmie, A. (2017). Beyond translation: adapting a performance-task-based assessment of critical thinking ability for use in Rwanda. *Assess. Eval. High. Educ.* 42, 673–689. doi: 10.1080/02602938.2016.1177484

Shavelson, R. J. (2010). *Measuring College Learning Responsibly: Accountability in a New Era*. Redwood City, CA: Stanford University Press.

Shavelson, R. J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., and Marino, J. P. (2019). Assessment of university students' critical thinking: Next generation performance assessment. *Int. J. Test.* 19, 337–362. doi: 10.1080/15305058.2018.1543309

Tremblay, K., Lalancette, D., and Roseveare, D. (2012). *Assessment of Higher Education Learning Outcomes AHELO. Feasibility Study Report. Volume 1: Design and Implementation*. Paris: OECD.

Tuononen, T., Hyytinen, H., Kleemola, K., Hailikari, T., Männikkö, I., and Toom, A. (2022). *Systematic review of learning generic skills in higher education - enhancing and impeding factors.*

Ursin, J., Hyytinen, H., and Silvennoinen, K. (eds) (2021). *Assessment of Undergraduate Students' Generic Skills in Finland: Findings of the Kappas! Project*. Helsinki: Ministry of Education and Culture.

van Someren, M. W., Barnard, Y. F., and Sandberg, J. A. C. (1994). *The Think Aloud Method: A Practical Guide to Modelling Cognitive Processes*. London: Academic Press.

van Widenfelt, B. M., Treffers, P. D. A., de Beurs, E., Siebelink, B. M., and Koudijs, E. (2005). Translation and cross-cultural adaptation of assessment instruments used in psychological research with children and families. *Clin. Child Family Psychol. Rev.* 8, 135–147. doi: 10.1007/s10567-005-4752-1

Zahner, D., and Ciolfi, A. (2018). "International comparison of a performance-based assessment in higher education," in *Assessment of Learning Outcomes in Higher Education: Cross-national Comparisons and Perspectives*, eds O. Zlatkin-Troitschanskaia, M. Toepper, H. A. Pant, C. Lautenbach, and C. Kuhn (Berlin: Springer International Publishing), 215–244. doi: 10.1007/978-3-319-74338-7

Zlatkin-Troitschanskaia, O., Shavelson, R. J., Schmidt, S., and Beck, K. (2019). On the complementarity of holistic and analytic approaches to performance assessment scoring. *Br. J. Educ. Psychol.* 89, 468–484. doi: 10.1111/bjep.12286

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.