DOI: https://doi.org/10.15366/jospoe2021.13.003
JOURNAL OF SUPRANATIONAL POLICIES OF EDUCATION, nº13, pp. 62-87

# INTERNATIONAL PERFORMANCE ASSESSMENT OF CRITICAL THINKING: FRAMEWORK FOR TRANSLATION AND ADAPTATION

# PRUEBAS INTERNACIONALES DE DESEMPEÑO PARA LA EVALUACIÓN DEL PENSAMIENTO CRÍTICO: MARCO PARA LA TRADUCCIÓN Y ADAPTACIÓN

Natalia Ronderos

Richard J. Shavelson

Doreen Holtsch

Olga Zlatkin-Troitschanskaia

Guillermo Solano-Flores

## ABSTRACT

Higher education institutions worldwide claim they impact students' learning outcomes within and across academic domains. Critical thinking (CT) is prominent among the intended outcomes (Braun et al., 2020). In this context, there is increasing interest in ecologically valid performance assessments (PAs) of CT that can be used internationally (Zlatkin-Troitschanskaia et al., 2018). While several studies have aimed to measure and compare CT skills across countries, this typically has been done using multiple-choice questions. Only a few studies involve adaptation of PAs. Their results point to the need for a more refined adaptation process (Braun et al., 2020), especially in terms of functionally equivalent adaptation and redesign. Based on a review of previous approaches related to test adaptation, with focus on the challenges of achieving cultural responsiveness, we propose a conceptual framework for adapting PAs of CT for international studies. The framework differentiates between two stages and three adaptation designs. The first stage involves a collaborative approach to the design of PAs of CT. The second stage offers three design alternatives which differ in their emphasis on linguistic considerations and cultural responsiveness. While this paper focuses on PAs of CT for higher education, it may be applicable to pre-college education.

**Key words**: Critical thinking; Performance assessment; Translation; Adaptation; International assessment.

## RESUMEN

En todo el mundo, las instituciones de educación superior afirman que influyen en el aprendizaje de los estudiantes dentro y a través de distintas áreas académicas. El pensamiento crítico destaca entre las áreas de más interés (Braun et al., 2020). En este contexto, existe un interés creciente en pruebas de desempeño de pensamiento crítico que sean ecológicamente válidas y que se pueden utilizar internacionalmente (Zlatkin-Troitschanskaia et al., 2018). Varios estudios internacionales han tenido como objetivo medir y comparar el pensamiento crítico entre países, pero empleando preguntas de opción múltiple. Pocos estudios involucran la traducción y adaptación de pruebas de desempeño y los resultados que han producido apuntan a la necesidad de mejorar el proceso de adaptación (Braun et al., 2020). Basados en una revisión crítica de enfoques y marcos previos relacionados con la adaptación de pruebas, y con el fin de superar las dificultades para logra una mayor sensibilidad cultural, proponemos un nuevo marco conceptual para la adaptación de pruebas de desempeño de pensamiento crítico para estudios internacionales. El marco propuesto distingue dos etapas y tres diseños de adaptación. La primera etapa presenta un enfoque colaborativo para

el diseño de las pruebas. La segunda ofrece tres opciones de diseño con diferentes grados de énfasis en aspectos lingüísticos y sensibilidad cultural. Aunque el artículo se enfoca a las pruebas de desempeño de pensamiento crítico para educación superior, se le puede aplicar en contextos preuniversitarios.

**Palabras clave**: Pensamiento crítico; Pruebas de desempeño; Traducción, Adaptación; Pruebas internacionales.

Date of receipt: 1 april 2021.
Date of acceptance: 9 june 2021.

## 1. INTRODUCTION

Higher education institutions worldwide claim they impact students' learning outcomes within and across academic domains. Critical thinking (CT) is prominent among outcomes (Braun et al., 2020; Liu et al., 2014). CT is justified both by professional-field requirements (O'Leary et al., 2020) and civic and citizenship engagement (Shavelson et al., 2019). Additionally, with internationalization and growing accountability demands, interest has increased in ecologically valid CT assessments that can be used nationally and internationally (Zlatkin-Troitschanskaia et al., 2018).

International assessments of learning outcomes have been used in elementary and secondary education in specific subject-areas (Wagemaker, 2010). Their use has grown exponentially over the last decades and gained momentum in the public debate on education (Carnoy, 2019). Countries participate in these assessments, amongst other reasons, as a way to provide evidence on which to formulate or evaluate policy, quantify their human capital competitiveness, and enhance curriculum and pedagogy (Addey & Sellar, 2017; Solano-Flores, 2019b). These same reasons have been used to justify the benefits of international assessments of CT in higher education (Tremblay, 2013). The aim is to evaluate higher education claims and compare CT outcomes between students and programs and across countries.

Given its importance, CT has to be carefully defined, as it provides the basis for its subsequent assessment (e.g., Liu et al., 2014). However, CT's definition is highly contentious (Braun et al., 2020; Lieu et al., 2014). The International Performance Assessment of Learning (iPAL) collaborative[1] defines CT as a multifaceted construct. CT involves conceptualizing, analyzing, synthesizing, evaluating, and applying information to solve a problem, decide on a course of action, find an answer to a given question, reach a conclusion, or some combination of these while avoiding judgmental biases. It also involves communicating the results of CT clearly and concisely (Shavelson et al., 2019).

CT assessments have been mostly multiple-choice (Liu et al., 2014). However, given iPALs definition of CT, it involves higher-order thinking abilities requiring multiple cognitive processes and dispositions—a multiple-choice format would result in construct underrepresentation (Braun et al., 2020). Similarly, a common critique of multiple-choice items is that they "make it infeasible to assess some of the more complex cognitive processes that correspond to ambitious curriculum aspirations (…) Such items are quite effective at measuring knowledge of fact, procedures, and concepts (…) there are limits to these formats." (Linn, 2002, p. 35) Performance assessments (PAs) provide an alternative to assess CT. They simulate, as closely as possible, a real-life situation. This verisimilitude results in more accurate construct representation and lower construct irrelevant variance (Braun et al., 2020). As such, PAs are the most suited assessment approach if one follows the iPAL definition (Braun et al., 2020; Hyytinen & Toom, 2019; Oser and Biedermann, 2020; Zlatkin-Troitschanskaia et al., 2018).

For international assessments, translation and adaptation are fundamental processes to ensure the validity of test result interpretations, achieve comparability of test scores, and guard fairness. Most translation and adaptation guidelines are either too broad, covering the complete range of educational tests (e.g., International Test Commission [ITC], 2017), or are purposely targeted for international subject-specific assessments that use prevalently multiple-choice items (e.g., PISA and TIMSS). Differences in PAs' format (e.g., nature and extent of the stimulus, the document library, the degree of contextualization that is required) and the construct of CT (situated in the practical use for decision-making) represent content and linguistic challenges that need to be carefully considered. The specificity of PAs of CT and their differences with subject-specific multiple-choice tests are so significant that the use of current translation and adaptation guidelines developed for multiple-choice formats is insufficient and potentially problematic.

[1] iPAL aims to collaboratively develop reliable and valid performance assessments of 21st century cross-disciplinary ("generic") and domain-specific skills that can be used by higher-education institutions nationally and cross-nationally to measure learning (Shavelson et al., 2018).

To date, only a few studies involve the translation and adaptation of PAs of CT: The Assessment of Higher Education Learning Outcomes (AHELO), CLA+, and iPAL. A significant shortcoming of AHELO performance assessments was the inadequate contextualization of the tasks to more than one country (the US in this case) (Tremblay et al., 2013; Shavelson et al., 2019). Although CLA+ and iPAL have attempted to address these challenges, the results have not yet been systematically analyzed, and a more refined adaptation process is required (Braun et al., 2020).

In this article, we focus on the research question: *How can PAs of CT be developed, translated and adapted for international studies, with particular attention to the challenges of achieving authenticity in the local context while, if desired, maintaining functional equivalence and thus international comparability*. More specifically, we ask: *What is a suitable framework for the adaptation of PAs of CT for use in international and national comparisons?* We build on earlier literature and iPALs experience to propose a conceptual framework for adapting PAs of CT for these purposes.

## 2. CRITICAL THINKING AND PERFORMANCE ASSESSMENTS

The definition of CT has been debated regarding its (i) universality, (ii) generalizability, and (iii) scope. Regarding the universality of CT and its relation to education, Beck, 2020, and Oser and Biedermann, 2020, offer interesting reflections on the differences in its meaning between the European and North American traditions (Beck, 2020; Oser and Biedermann, 2020) and caution that for international studies a common definition first needs to be established (Beck, 2020). Additionally, there is an ongoing debate regarding CTs domain-specificity or generalizability (Liu et al., 2014; Nagel et al., 2020, in Braun et al., 2020; Oser and Biedermann, 2020; Siegel, 2010). Finally, the scope of the definition varies, for example by including or excluding attitudinal aspects (Hytineen & Toom, 2019; Liu et al., 2014; Mihaildis & Thevenin, 2013; Siegel, 2010).

Oser and Biedermann (2020) address these disagreements and identify three different levels on which CT manifests itself: critical analysis, critical reflection, and critical alertness. The first level, critical analysis, requires domain-specific knowledge; the second, critical reflection, is generic and framed within individuals' societal responsibility; and the third is more attitudinal (Oser and Biedermann, 2020). iPAL takes an approach close to what Oser and Biedermann (2020) define as critical reflection and critical alertness.

Given iPAL's definition, multiple-choice questions might lead to construct underrepresentation. For CT, a criterion-sampling approach respects the construct, understands its complexity and conceives the whole as being more than the sum of its parts (Shavelson, 2011). With a criterion-sampling approach, criterion situations are sampled from real-world contexts and used to assess performance (McClelland, 1973). PAs, then, are built upon a criterion-sampling approach to assess CT.

PAs of CT introduce students to a situation based on a real-word event (storyline) and ask that they take a specific role (e.g., advise an official). Students are provided a document library that includes a variety of points of view and information sources (e.g., newspaper article, blogs, research reports, journal articles, webpages, among others) (Shavelson et al., 2018). This information varies in trustworthiness, relevance, and possible judgmental bias (Braun et al., 2020). They ask students to justify their decisions, recommendations, etc. given the information provided (Shavelson et al., 2018). These tasks are complex and do not offer a single solution path (Shavelson et al., 2018).

Consider as an example the PA, "Refugee crisis," developed within iPAL (Braun et al., 2020; Hyytinen & Toom, 2019): The storyline is about a fictitious country that is facing increasing demand for migrants' entry. The government has to decide whether to increase migration and reception centers for migrants, in light of claims of the relationship between migrants and increase in crime. The students are asked to enumerate the pros and cons of accepting more refugees, supporting them with evidence from the document library. Additionally, students are prompted to

elaborate and recommend a concrete course of action, backing up their recommendation with evidence from the documents.

There has been interest in applying PAs of CT like the PA "Refugee crisis" internationally. However, there have been few international studies in higher education of CT, using PAs. In all of them, adaptation has been a fundamental process to enhance their quality in measurement equivalence, the validity of score interpretations, and fairness. However, there are still shortcomings in this regard (see Section 4.1).

## 3. CONCEPTUAL FOUNDATIONS FOR TEST TRANSLATION AND ADAPTATION

### 3.1 DEFINITION OF TRANSLATION AND ADAPTATION

The literature often distinguishes between test translation and test adaptation (e.g., Ercikan & Por, 2020; ITC, 2017). At times both concepts converge (e.g., Solano-Flores et al., 2009). Therefore, we define both concepts. Translation refers to the creation of different language versions of a test that are linguistically equivalent. Adaptation refers to a broader process that, in addition to language, includes cultural considerations such as equivalence of the construct and familiarity with the item format (Berman et al., 2020; Ercikan & Por, 2020; van de Vijver & Poortinga, 2016). Adaptation has been generally preferred as it explicitly represents the complexities of different language versions across cultures (Ercikan & Por, 2020; Hambleton., 2005; ITC, 2017). However, translation is a very complex process. Translation inevitably involves culture, as two languages, for example, may express the same content in different culture-specific ways.

Simply put, languages are complex. They vary in their grammatical forms, word usage, and difficulty (Berman et al., 2020). Languages encode meaning in different ways (Solano-Flores, 2009). Since translation relies on the process of decoding and recoding meaning, it also requires the understanding of how meaning is shaped by culture (Solano-Flores, 2012)[2]. This richness and complexity of languages and the fact that "languages are far from having a word-to-word correspondence" (Solano-Flores et al., 2009, p. 80) means that all translation requires at least some degree of adaptation. This is evident in the fact that even if aiming at linguistic equivalence, two translated versions of a text will seldom be identical and at least some degree of variation is to be expected. The degree of variation in translations may depend on characteristics of the target language and its interplay with the intended meaning of a test item and the features of the source and target culture and population.

In recognition that translation inevitably involves adaptation, we use the term adaptation henceforth and follow the definition of adaptation established by the translation and adaptation guidelines (TAGs):

> Test adaptation refers to all of the activities including: deciding whether or not a test in a second language and culture could measure the same construct (…); selecting translators; choosing a design for evaluating the work of test translators (…); choosing any necessary accommodations; modifying the test format; conducting the translation; checking the equivalence (…) and conducting other necessary validity studies (ITC, 2017, p. 7).

From this definition, we emphasize that the adaptation process starts with reflections related to the equivalence of the construct across cultures, and ends with statistical and judgmental validity studies to verify the cognitive equivalence of the test versions. Our approach, anchored to the TAGs, considers both cultural and cognitive aspects of adaptation from beginning to end. We note

[2] Additional terms that are sometimes used to refer to adaptation are `localization´ (ITC, 2017; Solano-Flores, 2012) and `transfer´ (van de Vijver & Poortinga, 2016). In this article, we refer to these two terms as `adaptation´.

that the language may be the same but cultural difference between countries may vary in important ways (e.g., Holtsch et al., 2016), and that within a country and language, there might be relevant cultural differences.

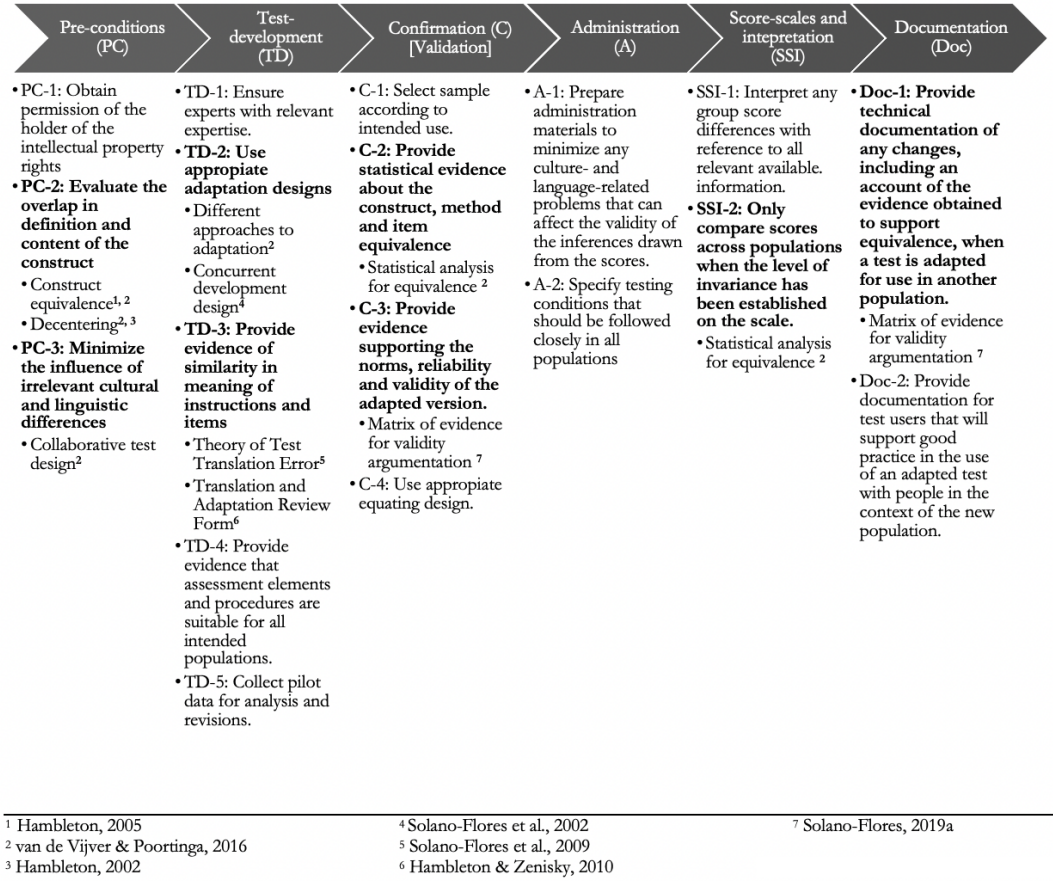
The conceptual foundations that guide our approach to test adaptation begin with recognizing assessment instruments as cultural products (Berman et al., 2020; Brown, 2016; Ercikan & Solano-Flores, 2016; Mislevy, 2018; Solano-Flores, 2011; Solano-Flores, 2019a). This recognition establishes that individuals interact with assessment according to their previous experience and their social, cultural and linguistic contexts (Miselvy, 2018). The relationship between assessment and the sociocultural and linguistic context has implications for cross-cultural assessments in the most fundamental concepts in educational measurement as noted above: (a) comparability and equivalence, (b) validity, and (c) equity and fairness.

### 3.2 GUIDELINES AND FRAMEWORKS FOR ADAPTATION[3]

Several methodological approaches and guidelines for assessment adaptation have been developed and refined over the last years, spurred on by rapid growth in the field (ITC, 2017). The most widely acknowledged approach to assessment adaptation are ITC's (2017) TAGs. The first edition of the TAGs was published in 1996, and there have been significant advances in the field since (ITC, 2017). The latest edition presents 18 guidelines, understood as essential practices to follow at the different stages of the adaptation process (ITC, 2017). These are organized into six sections: Pre-conditions (PC), Test-development (TD), Confirmation (C), Administration (A), Score-scales and interpretation (SSI), and Documentation (Doc). Figure 1 anchors the TAGs and other frameworks, and highlights in bold eight specific guidelines that we address in this section. They are considered essential for the adaptation of PAs of CT.

[3] Some of the most influential work on methods for adaptation include the ITC translation and adaptation guidelines (TAGs) (ITC, 2017); Hambleton's work in relation to the TAG's (Hambleton, 2002; Hambleton., 2005; Hambleton & Zenisky, 2010); van de Vijver's contributions regarding differential approaches to test adaptation and statistical methods to confirm equivalence (van de Vijver, 2016; van del Vijver & Poortinga, 2014, 2016); and Solano-Flores's analytical frameworks and detailed concurrent development approach (Solano-Flores et al., 2009; Solano-Flores, 2019a; Solano-Flores et al., 2002). These advances have benefited from and been beneficial for large-scale international educational surveys such as those of the IEA (e.g., TIMSS and PIRLS) and OECD (e.g., PISA).

Figure 1: TAGs and other approaches for translation and adaptation of tests




Source: TAGs (ITC, 2017). Figure by the authors.

The eight highlighted guidelines are: PC-2, PC-3, TD-2, TD-3, C-2, C3, SSI-2 and Doc 1. Pre-condition 2 (PC-2) and PC-3 relate to construct equivalence among cultures and aspects that facilitate or hinder the ensuing adaptation. In the Test Development (TD) stage, TD-2 is about choosing the corresponding subsequent adaptation process, and TD-3 is about the review of the quality of the adapted version. Confirmation stage C-2, C-3, and score scale and interpretation (SSI) SSI-2 and Documentation (Doc) Doc-1 are about appropriately gathering, documenting and using the equivalence and validity evidence for the adapted version of a tests.

Given the encompassing nature of TAGs in the development processes and recognizing the diversity of possible designs and approaches for each stage, we use them as anchors to present other complementary approaches from Hambleton, Solano-Flores, and van de Vijver. These are linked to specific guidelines. In Figure 1, these are referenced in each of the eight highlighted guidelines, and the corresponding reference is included in the footnote.

### 3.2.1 Pre-Conditions: overlap in the construct (PC-2) and minimizing the influence of construct-irrelevant cultural and linguistic specifics (PC-3)

Hambleton (2005), Hambleton (2002) and van de Vijver & Poortinga (2016) clarify the meaning of construct equivalence, and how to capture it in practice (PC-2). Hambleton (2005) argues that construct-equivalence is a pre-requisite for any adaptation study and considers its failure to do so as "one of the most serious errors in cross-language research." (Hambleton, 2005, p,7) Hambleton defines construct equivalence as "both conceptual/functional equivalence as well as equivalence in the way the construct measured by the test is operationalized in each language/cultural group" (Harkness, 1998 in Hambleton, 2005, p. 6). Hambleton (2005) recommends relying mainly on judgmental approaches to establish construct equivalence.

If construct equivalence cannot be established, then they recommend discontinuing the project or considering "decentering", "i.e., revising the definition of the construct to be equivalent in each language and cultural group" (Hambleton, 2002, p. 65). Van de Vivjer et al. (2004) set forth "retroactive" procedures for analyzing construct equivalence statistically after the test has been applied and data collected. We refer to construct equivalence when we discuss guideline C-2: statistical analysis of equivalence.

For minimizing construct-irrelevant variance due to linguistic and cultural specifics (PC-3), van de Vijver & Poortinga (2016) emphasize the benefits of collaborative development of the instruments amongst representatives of the diverse cultures. They thus recommend engaging stakeholders in test development and that members of all target cultures participate (van de Vijver & Poortinga, 2016). In IEA's and OECD's studies, for example, participating countries propose items for the test's item pool (Linn, 2002; van de Vijver & Poortinga, 2016).

### 3.2.2 Test development: adaptation design (TD-2) and evidence of similarity in meaning (TD-3)

Van de Vijver & Poortinga (2016) distinguished three approaches to adapting tests—adoption, adaptation, and assembly. Adoption relies on a precise translation and results in very similar formats and contents between the source and target versions of the test. Adaptation relies on retaining some items and modifying those that do not transfer well. Assembly encompasses the development of a new instrument for a target culture while keeping the construct from the source test. In assembly, both the content and format differ greatly in both versions while retaining the construct.

Adaptation designs, including those that could be categorized in any of the three approaches proposed by van de Vijver & Poortinga (2016), have prevalently been successive. That is, there is a source version of the test, developed in a source culture and language, that is later adapted to a

target culture and language. Even if stakeholders engage in the test development process (as suggested also by van de Vijver & Poortinga, 2016) there is still a source version of the test at the beginning, and then that version is adapted. In contrast to this approach, Solano-Flores et al. (2002) propose a concurrent development process using item shells (Solano-Flores et al., 2002).

Solano-Flores et al. (2002), consider the concurrent development design as the only approach to achieve comparable and equitable evidence as "both languages, and their speakers are given the same opportunities to influence the process of assessment development" (Solano-Flores et al., 1999 in Solano-Flores et al., 2002, p. 109). In this process, item shells are used as blueprints for the assessment, which is simultaneously designed in the different languages/cultures. The test-development process follows the same steps in both languages/cultures and thus equitably considers them (Solano-Flores et al., 2002).

Evaluations of the adaptation that attempts to maintain the construct are essential to ensure (functional) equivalence. Solano-Flores et al. (2009), and Hambleton & Zenisky (2010) provide useful tools. Solano-Flores et al. (2009) offer The Theory of Test-Translation Error (TTTE). TTTE defines error as multidimensional and inevitable (Solano-Flores et al., 2009). Translated items or tasks are reviewed, focusing on identifying errors, and they are deemed acceptable or objectionable according to the frequency and severity of error (Solano-Flores et al., 2009). Hambleton & Zenisky (2010) provide a review protocol to aid the standardization of the judgmental reviews (Hambleton & Zenisky, 2010).

### 3.2.3 Confirmation (Validation), administration, score scales interpretation and documentation

The confirmation (validation) stage of the adaptation process involves conceptual and empirical examination of the "quality" of the assessment: the relevant conceptual and statistical evidence about the construct, method, and item equivalence. Van de Vijver & Poortinga's (2004, 2016) approaches in this regard aim at defining the different levels of equivalence that are desirable and should be explored, as well as proposing pragmatic solutions to the analysis of each level and its feasibility. Although they emphasize the statistical component, they also recognize the role of judgmental considerations to explore bias and confirm equivalence. Van de Vijver & Poortinga (2004, 2016) establish four levels of equivalence: (i) conceptual, (ii) structural, (iii) metric and, (iv) full-score or scalar (van de Vijver & Poortinga, 2016). The conceptual level refers to the equivalence of the meaning of the construct in both cultures; the structural, or functional, level refers to similar factor structures of the items and interrelations between them; the metric level refers to measurement unit equivalence, and the level to full-score or scalar equivalence. The latter is the most stringent and implies that scores can be compared at face value (van de Vijver & Poortinga, 2016). The authors note that scalar equivalence is impractical, and thus although theoretically desirable, it cannot be the expectation for cross-cultural studies (van de Vijver & Poortinga, 2016). They thus establish different alternatives (van de Vijver & Poortinga, 2016). Unless partial equivalence or invariance in the sense developed by Byrne et al. (1989) has been established, there should not be comparisons of test-scores. These need to be interpreted carefully depending on the equivalence level that has been established.

It is essential that for the interpretation the administration conditions are taken into account. According to the TAGs (2017), some issues in administration that can compromise fairness are the clarity of instructions, motivation and scoring of items. To address this difficulty standardization of administration is essential, along with flexibility to introduce accommodations when necessary (ITC, 2017).

Finally, the confirmation (validation) and documentation phases focus on communication and documentation. Solano-Flores (2019a) offers the Matrix of Evidence for Validity Argumentation, a method for systematically capturing and integrating validity evidence of the entire assessment

process. At its core is the explicit recognition of the ubiquity of cultural issues in all stages of assessment and a call for a more systematic, conceptual, and operational approach to cultural responsiveness in cross-cultural assessments (Solano-Flores, 2019a).

#### 3.2.4 Contribution and insufficiency of the tags for the adaptation of performance assessments of critical thinking

The TAGs and other relevant approaches have helped international assessment agencies such as the OECD and IEA improve their adaptation guidelines and procedures (Linn, 2004; Wagemaker, 2010). International assessments have applied these guidelines by developing specific and more detailed protocols, which mainly focus on subject-specific, multiple-choice tests.
In this context, Ercikan & Por (2020) note: "The current guidelines for test adaptations have yet to consider the possibilities and limitations of the new assessment types, partly due to the dearth of research studies." (p, 219) As such, the previously outlined approaches are necessary but not sufficient for the adaptation of PAs of CT. The TAGs should be applied to adaptation of PAs of CT that aim at comparability (Shavelson et al., 2010). Additionally, all of van de Vijver's approaches presented here can be transferred to the adaptation of PAs of CT. Moreover, Solano-Flores's concurrent development process (Solano-Flores et al., 2002), TTTE (Solano-Flores et al., 2009), and Matrix for Validity argumentation (2019a) are useful conceptual and operational tools. However, for the test-development stage outlined in the TAGs, more specificity is required for PAs of CT to address the challenges as described in Section 4.1.

### 3.3. FUNDAMENTAL CONCEPTS FOR ADAPTATION

#### 3.3.1 Comparability and (functional) Equivalence

When translating and adapting PAs of CT, (functional) equivalence should be taken into account to enable comparability. In the TAGs, the concept of comparability is the aim of any test adaptation (ITC, 2017), and equivalence is the means to achieve this goal.

Comparability is the extent to which:

> … students' scores can be validly compared … even if those scores come from measurements taken at different times, in different places, or using variations in assessment content and procedures. Ideally, users could be assured that students with the same score possessed the same level of proficiency… (Berman et al., 2020, p. 14)

Tests are comparable when they have the same: (1) purpose, design, and interpretation; (2) content and construct domain; (3) measurement properties; (4) administration conditions; and (5) student background factors, e.g., linguistic, and sociocultural (Berman et al., 2020, p. 4). In cross-cultural studies, the most obvious divergence case is due to instruments differing in language versions (Bennet, 2020) and students' differing in background. Particularly for international assessments, Ercikan & Por (2020) consider the following three criteria to support comparability of international tests: "(1) the assessments are tapping the knowledge and skills we are interested in assessing, (2) the constructs being assessed are comparable for different sociocultural groups, and (3) the scores are comparable across languages and cultures (Ercikan & Lyons-Thomas, 2013)." (p. 206)
Both Ercikan & Por (2020) and van de Vijver & Poortinga (2004, 2016) stress the importance of equivalence to enable test-score comparisons between groups. Equivalence is a property of the measurement and has different levels. Van de Vijver & Poortinga (2016) propose equivalence levels that were previously mentioned: construct, structural, metric, and scalar. The TAGs acknowledge this typology and also suggest considering method and item equivalence.

The most basic tenet of equivalence is construct equivalence, which cannot be taken for granted in cross-cultural comparisons, as "sometimes the differences among cultures or individuals' backgrounds prove too profound to proceed as if we are measuring the same construct with different forms of the same assessment" (Mislevy, 2018, p. 219). Once construct equivalence has been judgmentally established, the adaptation has been made, and an application has occurred, additional judgmental and statistical analysis is needed to confirm both construct and other equivalence types (van de Vijver & Poortinga, 2016; ITC, 2017).[4] Due to the close relationship between equivalence and comparability, we consider these two concepts to be essential for all international comparative studies.

### 3.3.2 Validity

Validity is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (American Educational Research Association [AERA] et al., 2014, p. 11). The impact of culture on the validity of assessment interpretation has long been recognized as a challenge, driving test designers to propose culture-free, culture-reduced, or culture-fair assessments (van de Vijver, 2016). Solano-Flores (2011) proposes consideration of culture to ensure validity, and the Matrix of Evidence for Validity Argumentation (2019) puts the concept at the center of cross-cultural assessments. He states:

> when students from different cultural backgrounds are assessed with the same instrument, fairness [and] the validity of the interpretations of test scores becomes (sic) an issue … In this case, test score differences are attributable, at least to some extent, to cultural differences rather than differences on the target knowledge or skills (AERA et al., 2014) (Solano-Flores, 2019a, p. 2).

Validity can be generally compromised due to construct-irrelevant variance (Sierci & O'Riordan, 2020) when in cross-cultural assessments, one language/culture group interacts with the assessment differently than another. Construct-irrelevant variance could be caused by adaptation error due to (1) cultural/language differences; and (2) technical issues, designs, and methods (Hambleton, 2005). Construct-irrelevant variance can lead to bias (van de Vijver, 2016) and validity, equivalence, and fairness issues. Additionally, validity can also be compromised by construct underrepresentation. This would be the case if in "decentering" the construct, as proposed by Hambleton (2002) part of it is left aside.

### 3.3.3. Fairness

The Standards for Educational and Psychological Testing (hereafter: Standards) consider fairness as a fundamental validity issue (AERA et al., 2014). The Standards explicitly relate fairness to equivalence and comparability. Fairness is defined as reflecting the same construct and permitting equal interpretation of tests scores for all individuals in a way that "does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct." (AERA et al., 2014, p. 50) The Standards, then, establish that testing should be responsive to diverse groups when their characteristics can compromise validity (AERA et al., 2014). The standards present four views of fairness: (i) treatment during the assessment process; (ii) lack of bias; (iii) access to the constructs measured; and (iv) validity of individual score interpretation (AERA et al., 2014).

Out of concern about the assessment process, both equity and fairness have led to standardization and to the idea of test accommodations or adaptations. Standardization is considered a way to

[4] Due to word count limitations, we cannot go into further detail regarding equivalence; for further details on equivalence and functional equivalence, see van de Vijver & Poortinga (2004) and Braun (2006).

ensure equal conditions for all by reducing variation on irrelevant aspects to the construct and thus possible construct irrelevant variance (AERA et al., 2014; Mislevy, 2018). However, when dealing with diverse populations, standardization might act against its purpose of leveling the playing field (Sireci et al., 2020). In such cases, "greater comparability of test scores may be attained if standardized procedures are changed to address the needs of specific groups or individuals" (AERA et al., 2014, p. 51). The idea that greater fairness can be obtained through flexibility is what Mislevy (2018) proposes as a conditional sense of fairness: in certain cases, surface conditions of the test and administration can be varied to strengthen fairness and thus equivalence and validity.
The concept of a conditional sense of fairness leads to a particular view of adaptation for diverse cultural groups. By introducing templates[5] it permits variation on surface (construct-irrelevant) aspects of the test while maintaining the construct (Mislevy, 2018). This view of adaptation for diverse groups is highly culturally responsive. It resembles the assembly option proposed by van de Vijver & Poortinga (2004 and 2016) and the concurrent process recommended by Solano-Flores et al. (2002).

### 3.4 CONTINUUM OF (FUNCTIONAL) EQUIVALENCE TRADE-OFF

The influence of language and culture on test-score interpretations' validity and fairness concerns the entire assessment process (Ercikan & Solano-Flores, 2016; Solano-Flores, 2019a). This complexity poses dilemmas and trade-offs between two ends of a continuum that aim to preserve the underlying construct and minimize any potential sources of construct-irrelevant variance that might influence differentially examinees' performance on tests across cultural or linguistic groups. The approaches differ on how they propose to reach equivalence. Berman et al. (2020) pose a major question: "how much and what types of variation in assessment content and procedures can be allowed, while still maintaining comparability across jurisdictions and student populations" (p. 19). We propose that cultural responsiveness is a requirement for comparability, and reframe the question: "In light of each particular test, how much and what types of variation is relevant to maintaining comparability?"
On the one hand, the argument would be to reach equivalence and comparability by aiming for maximal linguistic similarity between versions of the test (with similar content and format). This could imply minimal focus on other cultural aspects that would require changes in the format or content. This approach is achieved by precise translation focusing mainly on linguistics. On the other hand, the argument would be that equivalence, and thus comparability, arises from maximal cultural-responsiveness through the test's variation of surface-features. Any adaptation that aims at equivalence is a balancing act between a myriad of possibilities. Each approach represents a different choice as to the degree of loyalty to the source version of the test versus the degree of variation intended to maintain the test's meaningfulness for the target population.
The recognition of this continuum or balancing act is especially relevant for PAs of CT, due to the construct measured and the assessment format. CT is highly contextualized and often practical. Similarly, the definition of PAs of CT as criterion-sampling instruments from real-world situations, often in the social and civic domains, makes verisimilitude and the demand for cultural responsiveness, a very stringent criterion for the adaptation process. The bottom part of Figure 2, included in Section 4.2, illustrates this continuum along with the framework that we propose for adapting PAs of CT.

## 4. ADAPTATION OF PERFORMANCE ASSESSMENTS OF CRITICAL THINKING

### 4.1 CHALLENGES IN ADAPTATION

[5] Item templates serve as item shells, blueprints or frameworks (Mislevy, 2018).

Adaptation procedures should be highly responsive to the construct assessed and the format of the assessment. PAs of CT differ from multiple-choice and short-answer tests in multiple ways that affect adaptation. These differences include: (a) contextualization (decontextualized or briefly contextualized vs. deeply contextualized), (b) nature and extent of the stimulus (e.g., the short stem of a multiple-choice question vs. a storyline), (c) document library that can include different media (e.g., tweets, blogs, academic papers, etc.), (d) nature of the response (e.g., fill in bubble vs. open and extended written response), and scoring (e.g., dichotomous v. partial credit scores. These elements make the adaptation process of PAs of CT particularly challenging, likely making it the most complex challenge for adaptation in international assessments.

With respect to contextualization, PAs of CT emulate real-life, complicated situations, like the ones that college graduates are likely to encounter and that require their CT (Shavelson et al., 2018). PAs of CT are situated in everyday life challenges that arise in social, civic, economic, environmental, health, family, political, etc. contexts (Braun et al., 2020; Shavelson et al., 2019). Consequently, the "stimulus" or the "storyline" is far more complex than in multiple-choice or short-answer questions (Oser and Biedermann, 2020). Similarly, the documents included in the document library must have a high degree of verisimilitude. The fact that PAs of CT place students in real-life situations, highly dependent on sociocultural and linguistic patterns, makes adaptation particularly challenging.

Regarding the open and extended nature of the response, along with scoring, there is evidence that open response tasks can be more challenging for the adaptation process (for an illustration, Mislevy, 2018, p. 223). Open response tasks might lead to unpredicted responses, need a grading rubric, and often try to replicate more closely real-life situations (compared to choosing from existing alternatives as in multiple-choice questions). Students are free to construct answers that may vary, for example, in length, complexity, justifiability, and cultural embeddedness. Adaptation is also made more complex due to the nature of scoring. While multiple-choice tests can be automatically scored, PAs require raters scoring using rubrics. Additionally, to add to their complexity, as in real life, the challenge might have multiple solutions of varying justifiability on evidentiary and ethical grounds. Scoring is thus much more complex than with multiple-choice items both due to the characteristics of the open-ended response and raters' training to score those responses reliably. Adaptation of scoring materials also needs to be addressed and carefully reviewed as the equivalence of the assessment and, thus, comparability of scores, might be highly dependent on it.

Moreover, in cross-cultural assessment, familiarity with the assessment context plays an important role. PAs are typically an unfamiliar assessment context (Tremblay et al., 2013). Unfamiliar contexts impose higher cognitive loads than familiar contexts (Schendel & Tolmie, 2017) thus potentially increasing construct irrelevance variance and impairing students' abilities to perform effectively on a test (Solano-Flores et al., 2014). To this end, PAs should be accompanied by material to familiarize students with the nature of the PA and the expected response. For comparability of the interpretation, all cultural groups must be equally familiar with the format.

Perhaps due to these complexities, few PAs of CT have seldom been adapted and used internationally. The exceptions are: AHELO, CLA+, and iPAL. AHELO's PAs were critiqued as being: (a) inadequately contextualized (Shavelson et al., 2019) or "excessively 'American' in an international context" (Tremblay et al., 2013, p. 169), (b) of questionable content validity (Tremblay et al., 2013), and (c) reliable for only half of the countries (Tremblay et al., 2013). Generally, CLA+ studies have followed a similar approach to AHELO in favoring linguistic comparability. As with AHELO, these tasks have been developed in the United States and subsequently adapted to other countries. CLA+ international studies have encountered similar contextualization and authenticity problems to AHELO's (see, e.g., the baseball PT case in Zlatkin-Troitschanskaia et al., 2018). Those findings might partly be the result of taking PAs "off the shelf", in what has sometimes been

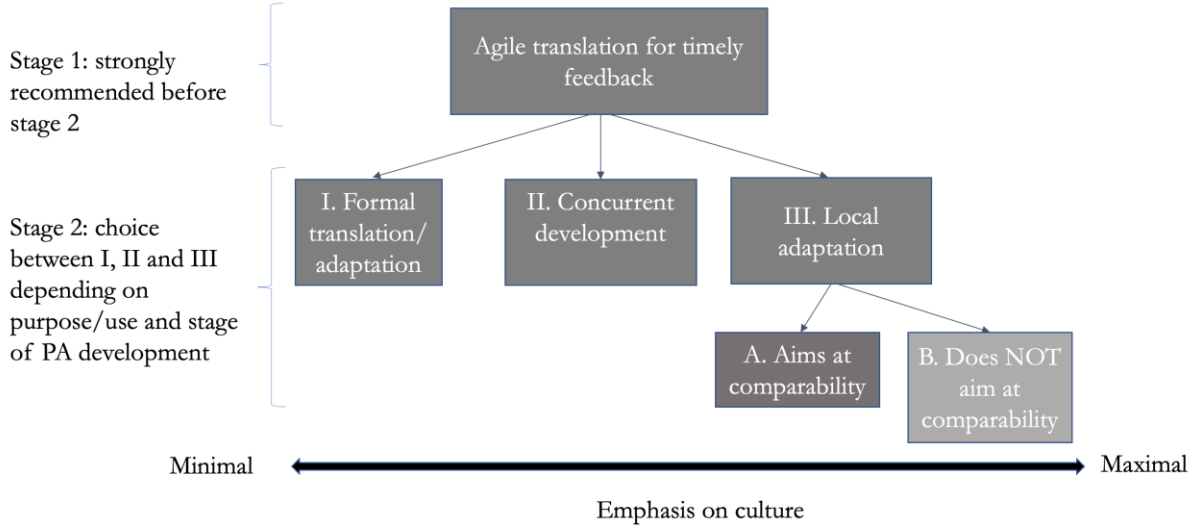
a cost-cutting decision, rather than creating these assessments with an international team of test developers, as iPAL suggests.

This section has established that PAs of CT present more pressing challenges for adaptation than other item formats. These challenges are evident in the shortcomings of previous international studies of CT through PAs. However, the approaches on which these studies are based and practical experience gained have not been systematically integrated into a coherent framework. In the following section, we present our proposal for integration. It incorporates previous experience from international studies of CT through PAs, and the various approaches recommended for adaptation.

### 4.2. FRAMEWORK FOR ADAPTING PERFORMANCE ASSESSMENTS OF CRITICAL THINKING

Due to the importance of the adaptation process in striving for comparability and equivalence, the validity of test score interpretations, and fairness (Ercikan & Por., 2020), it is essential to have a special framework for adapting PAs of CT. Our framework has two stages and three design choices for adapting PAs of CT (Figure 3). The two stages refer to (1) PA development incorporating timely international feedback and (2) choice of adaptation design. Stage 1 relies on an informal, rapid, "Google" (or other software) translation if not all participants fluently speak the language in which the PA is being developed. For stage 2, we differentiate between three adaptation designs choices that should be considered based on the aim of each particular study: (I) Formal adaptation (Shavelson et al., 2010; Solano-Flores et al., 2010); (II) Concurrent design (Solano-Flores et al., 2002); and (III) Local contextualized adaptation (see Schendel & Tolmie, 2017). Regardless of the choice of design, in all cases, formal interaction between teams of the source PA and the target PA is highly recommended, starting in stage 1 and throughout the whole process. Figure 2 illustrates the framework. Next, we describe each stage and design in more detail. Brief examples serve to illustrate some elements.

Figure 2: Framework for adapting PAs of CT




#### 4.2.1 Stage 1

The intent of stage 1, informal, rapid translation, is to promote a collaborative, cross-national development process. English could serve as a common language to enable this collaboration between participants who might not speak the same language as that of the PA being designed. We

suggest providing an English quick translation of a PA so that all test developers can comment. This is iPALs common practice. We consider it a "universal" stage.

Informal, rapid, automatic translation is intended to start the formal interaction with an international audience in the early stages of the development process. This step allows and invites the interaction and feedback of diverse cultural groups and responds to the general agreement regarding the fact that the "quality of adaptation is optimized when assessments in the source language are developed with the test adaptation goal in mind." (Ercikan & Por, 2020, p, 216)

In stage 1, once a rough storyline or a first draft of the storyline, documents and questions posed to students have been completed, they are translated using automated translation software or machine translation.

As the quality of automatic translation with engines such as Google Translate has improved, the number of scholarly evaluations has increased in the last decade (e.g., Aiken & Balan, 2011, Aiken, 2019; Taira et al., 2021). Moreover, the reception of automatic translation by users and researchers has evolved from an initial straight rejection to examining ways in which automatic translation can be used in combination with human translation and review in ways that optimize resources and accuracy (e.g., Stoltz, n.d.). Generally, there seems to be an agreement on the improvement in the accuracy of these translations (Aiken, 2019; Taira et al., 2021, Stoltz, n.d.). Recent studies have found high accuracy rates for certain language combinations (e.g., English and Spanish) (Aiken, 2019; Khoong et al. in Taira et al., 2021; Taira et al., 2021) but they also highlight high variability between languages, especially between Western and Asian languages (Aiken, 2019; Aiken & Balan, 2011; Taira et al., 2021). Informal sources of evaluation also point out to variability depending on the content or domain (Stoltz, n.d.).

Despite the improvements that have taken place, we are very far from a point in which we need not worry about accuracy of automatic translation. Thus, automatic translation can serve just as a starting point for the PA design team to work from by editing an initial automatic translation. By early international interactions on the test development process, diverse social, cultural, and linguistic patterns are considered from the beginning. Informal, rapid translation was used, for example, in adapting an original Finnish Migrants PA of CT to the Colombian context (Braun et al., 2020; Hyytinen & Toom, 2019). This experience taught valuable lessons to the Finnish, Colombian, and broader iPAL design and adaptation teams. During the conversations that followed this initial translation, what was learned led to adjustments in the original Finnish version of the PA, and to important adaptations of the Colombian version. For example, those conversations led to adjustments in a document that was intentionally included as irrelevant in the Finnish version, and became fundamental for assessing quantitative reasoning in the Colombian PA.

### 4.2.2 Design choices for Stage 2

For stage 2 we have identified three test-adaptation design choices to consider: (I) Formal adaptation, (II) Concurrent development, and (III) Local adaptation. The goal of formal adaptation (Design I) is to provide a PA that measures the same construct across multiple countries. There is an emphasis on linguistics, and as such, there are various steps including several independent translators. The aim of concurrent development (Design II) is to ensure equitable contributions and design process in all target languages and cultures (Solano-Flores et al., 2002). The aim of local adaptation (Design III) is to have a culturally responsive instrument valid for local use and comparisons within cultural contexts. Similar to the concurrent development process, this approach uses an existing PA as an "item shell" (e.g., Solano-Flores et al., 2002) while the storyline, documents, and scoring are locally adapted and can be changed from the original. In the local adaptation, there might be an intent to maintain the same construct as measured with the original PA (Design III-A) or not (Design III-B), depending on whether there is a comparative purpose in

the study. When used for comparative purposes, this approach aims to allow for variation on surface features while maintaining the construct, as proposed by Mislevy (2018) regarding the conditional sense of fairness.

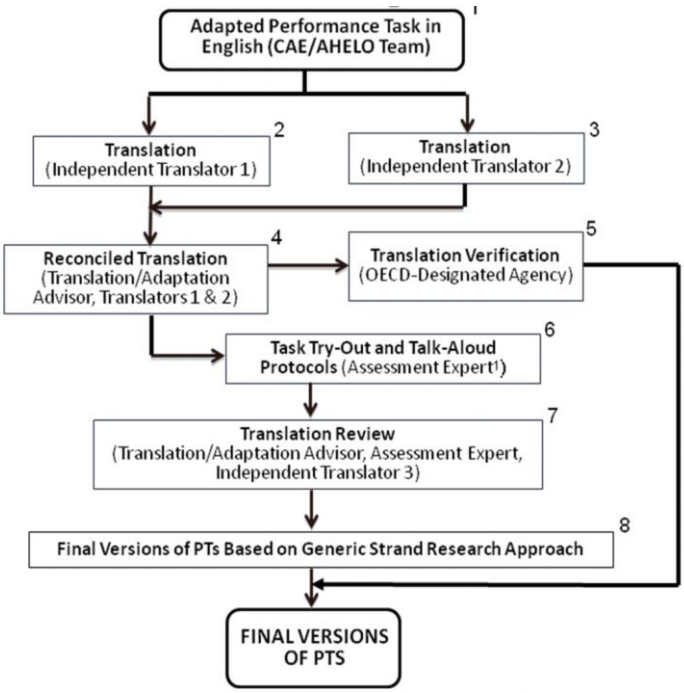
#### 4.2.2.1 Design I: Formal adaptation

The formal approach to adaptation builds on and integrates the TAGs (ITC, 2017), adjusting them to CTs PAs. This condition aims at assuring comparability by privileging content and format similarity through translation equivalence or linguistic similarity. However, there is some flexibility to incorporate adaptations necessary for construct equivalence in the target versions. We do not propose a specific process or flowchart for the formal condition. We recognize that the adaptation process' flexibility is vital to ensure its efficacy in different social and cultural settings (Zhao & Solano-Flores., 2020).

Whatever specific process for the formal adaptation is chosen, the methodology must acknowledge the adaptation process' complexity and thus proceed accordingly to what has been outlined in the TAGs (ITC, 2017) and TTTE (Solano-Flores et al., 2009). Any formal process should involve: certified and experienced translators, multiple translations done independently; a reconciled translation; revisions of the translation, and; a validation stage with talk-aloud protocols or cognitive laboratories.

As an exemplary implementation of Formal Design, we consider the AHELO guidelines (Shavelson et al., 2010; Solano-Flores et al., 2010). These guidelines provide an integrative conceptual framework that follows the TAGs (ITC, 2017) and TTTE (Solano-Flores et al., 2009). For reasons noted above, and in contrast to AHELO where participating countries were not involved in the initial PA development, we strongly recommend taking this approach only after Stage 1 (early international feedback).

Figure 3: AHELO's full translation procedure used for primary documents-PA and scoring rubrics



Source: Solano-Flores et al. (2010).

The full-translation process is illustrated in Figure 3. The PA origin was English (step 1 in Figure). This task was then translated into the target language by two independent translators (steps 2 and 3) and went through a reconciliation translation stage (step 4), with both teams participating, along with a translation/adaptation advisor. This stage's product was a reconciled translation that resulted from comparing and contrasting the two previous translations. This reconciled translation was then verified by an external, OECD-designated agency (step 5). Simultaneously, the resulting reconciled task went through a validation process that included task try-out and talk-aloud protocols (step 6). With the results of the validation, an additional translation review was done to ensure that the construct was maintained, the difficulty level was equal to the one of the original PA, and that students were interpreting the PA as expected (step 7 and 8). In the review stage, there were three participants: the translation/adaptation advisor who worked in the reconciliation, the assessment expert who carried the validations, and a third translator. With this process and incorporating the translation verification results done by the OECD-designated agency, the adaptation ended. The product was the final version of the PA in the target language (Solano-Flores et al., 2010).

#### 4.2.2.2 Design II: Concurrent development

Solano-Flores et al.'s (2002) experience and proposal are based on a context that aimed at fairly and equitably evaluating diverse linguistic and cultural groups within a nation, English language learners in the United States. However, we believe that this is an alternative design worth exploring, as it would be the direct continuation of stage 1: collaborative design through automatic translation. This condition would ultimately mirror what occurs with other assessment formats and international studies, where the item pools are drawn from the proposals of the participating countries (Linn, 2004; van de Vijver & Poortinga, 2016). The concurrent development process has been applied to multilingual studies (for example see Rogers et al., 2003) and supported with evidence on validity (Rogers et al., 2011), although not in the field of higher education, or PAs of CT.

#### 4.2.2.3 Design III: Local adaptation

The local approach uses assessment blueprints based on a source PA and its corresponding assessment framework. The approach has two variants: (A) one that aims to maintain the construct unaltered and achieve comparability of the target PA with the source PA, and (B) a more liberal adaptation of the source PA to local needs without any intention of comparability.

##### 4.2.2.3.1 Design III-A: Local adaptation for comparability

This design's goal is to keep fidelity to the source PA. We think of this design as replicating the concurrent development process' main elements (Solano-Flores et al., 2002) despite its successive nature. All elements in the assessment framework and source PA need to be maintained, and a judgmental process that assures its equivalence is necessary. The elements that must be equivalent in terms of the construct and assessment framework are the: (i) principal aspects of the storyline, (ii) questions asked of students, (iii) number of documents, (iv) nature of each document, (v) computer platform and interface, (vi) test application procedures (including material aimed at attuning and familiarizing students with PAs of CT), and (viii) scoring system. By the nature of each document, we refer to its purpose in the PA according to the iPAL assessment framework, presented in Braun et al. (2020): the perspective that it represents on the issue at hand (for or against, the aspects or considerations it privileges), the kind of document (newspaper article,

academic journal article, blog, scientific report, etc.), the nature of the information it has (qualitative information, video, text, quantitative data, etc.), and its degree of relevance, trustworthiness and potential bias.

There are no studies in the literature that document the local condition that aims for comparability. Zlatkin-Troitschanskaia et al. (2018) present an experience as part of a CLA+ international study. They followed this condition adapting a PA originally situated in baseball to soccer, a more familiar sport in Germany. However, the resulting PA's equivalence with the source PA has not been reported so far (Zlatkin-Troitschanskaia et al., 2018).

#### 4.2.2.3.2 Design III-B: Local adaptation with no aims of comparability

This design represents a more liberal adaptation, and thus the PA elements can be more freely changed. The original PA is used mostly as a reference and inspiration. Some of its features might be kept while others might be eliminated, and new elements might be added for verisimilitude. This condition aims at enabling local comparisons within universities of a country or even a region. As an illustration, consider the Rwandan CLA adaptation. The project sought to evaluate CT in Rwanda and did not have enough resources to design a new instrument (Schendel & Tolmie, 2017). The CLA's Crime PA (drugs and crime) was chosen to assess CT, but the evaluation team judged that the specific tasks proposed were unfamiliar to Rwandan students (Schendel & Tolmie, 2017). The project used the CLA Crime PA framework but replaced the drug-crime topic with two more pressing issues in Rwanda: road accidents and malaria (Schendel & Tolmie, 2017). Once the tasks had been adapted, experts' judgment, as well as student think-aloud protocols and field-testing, were used to gather validity evidence (Schendel & Tolmie, 2017). The CLA scoring method was adapted using fewer criteria and a different scoring rubric (Schendel & Tolmie, 2017). In concluding reflections, the evaluators defined this approach as "cultural adaptation", and reinforced its worth:

> Our experience strongly recommends the use of a cultural adaptation method when seeking to assess CT in a new cultural context. If the original version of the CLA had been used in the Rwanda study, it is clear the validity of the study results would have suffered substantially (…) the unfamiliar content of the CLA performance tasks would likely have introduced a significant amount of construct-irrelevant variance into the scoring distribution (Schendel & Tolmie, 2017, p. 685).

One question regarding the local condition's value with no aims at comparability might be a reason for adapting an existing PA of CT (Solano-Flores' blueprint) instead of developing a new one. We do not necessarily advise this; we just recognize that local adaptation is possible and could provide benefits compared to new PA development. Some of the benefits could be related to resources and efficiency (Schendel & Tolmie, 2017), while others relate to strengthening existing instruments, literature, and research (van de Vijver & Poortinga, 2016).

## 5. DISCUSSION AND CONCLUSIONS

This paper aimed to provide a framework for adapting PAs of CT to those who embark on international studies of CT using PAs. First, we noted that CT is considered to be one of the most prevalent learning outcomes in higher education. We proposed that it can be assessed by a criterion-sampling approach through PAs that authentically simulate situations that college graduates might encounter in and across multiple domains. Then, we explored the existing guidelines for adaptation, as well as other significant approaches. All approaches lead to recommendations regarding the participation of international, interdisciplinary teams, the use of both judgmental and statistical approaches, the recognition of the complexity of the adaptation and validation process, and the importance of considering the role of cultural specifics in these

processes, including valid interpretations, fairness, and equivalence if comparisons are intended in the respective study. The existing approaches have so far focused on multiple-choice items and applied in pre-college education. Consequently, we presented a framework that builds on them and adds specificity for PAs of CT for higher education. This framework for adapting PAs of CT based on the assumption regarding construct equivalence in all cultural and linguistic groups tested.

Our particular recommendations to those who embark on the adaptation of PAs of CT are to start developing PAs by incorporating early international feedback (stage 1) and then use this framework to decide what adaptation design they should use in their process (stage 2). Stage 1 utilizes online translation apps to permit collaborative development with representatives from diverse cultural and language backgrounds. Although machine translation is controversial, we recommend it as the fastest way to promote timely international collaboration, when the PA is being developed in a language in which not all the group members are fluent. We note, however, that the revision of the resulting translation, done by the test-design team is essential in this stage. This collaborative approach in test development for cross-cultural studies is highly recommended (ITC, 2017; Solano-Flores et al., 2002; van de Vijver & Poortinga, 2016) and facilitates further adaptation. This stage could be perceived as evident and thus be disregarded; however, we strongly recommend its pursuit. Many of the limitations of past international CT studies using PAs could be attributed to this missing step.

In stage 2, the formal adaptation process (Design I) relies mainly on the use of multiple translators and a reconciliation (Solano-Flores et al., 2010; Zlatkin-Troitschanskaia et al., 2018). It entails a rigorous translation approach and translation review (Solano-Flores et al., 2010), as well as gathering evidence on response process validity (Solano-Flores et al., 2010; Zlatkin-Troitschanskaia et al., 2018). This approach is suitable for international comparative studies and results in PAs similar in their linguistic characteristics, content, and structure. The main advantage of this design is the potential for (functional) equivalence and thus comparability. In fact, this is the only design with published studies that aim at international comparisons (e.g., Tremblay et al., 2012, Zahner & Ciolfi, 2018). There are two potential issues with this design: the resources required (time, staff and expertise involved), and its limited capacity for cultural responsiveness.

The concurrent development process (Solano-Flores et al., 2002) (Design II) proposes simultaneous development of PAs of CT with representatives of participating countries. It is intended to serve international comparative studies where simultaneous design settings can be pursued. It deepens collaboration and is characterized by its quest for equity and cultural responsiveness (Solano-Flores et al., 2002). The main limitation of this design is that it has not yet been used for PAs of CT, or in higher education. We propose it as a possibility and acknowledge the need for further research.

Similarly, the local approach (Design III) prioritizes cultural responsiveness and rests heavily on cultural validity concepts (Solano-Flores, 2019a). It differs from the concurrent process in that it is not necessarily done simultaneously. This approach could aim for comparability (Design III-A, for an example, see Zlatkin-Troitschanskaia et al., 2018) or be used to adapt existing PAs of CT for local use (Design III-B), on the bases of resource efficiency (Schendel & Tolmie, 2017) and further research into existing instruments (van de Vijver & Por., 2016). With Design III-A the main limitation, as with Design II, is the lack of research on the degree of equivalence that can be achieved. Although there is some past experience with this design, the final results have not been published (Zlatkin-Troitschanskaia et al., 2018). Both designs III-A and III-B share the same strength with design II, in terms of the emphasis placed on cultural responsiveness and thus local validity.

We recognize that our framework's implementation inevitably poses unforeseen challenges in the adaptation, scoring, validation and reporting of results. These challenges are largely dependent on context and available resources. We thus suggest that it is seen as an analytical tool and departure from establishing step-by-step procedures. We advise that any team that embarks on this endeavor

does so with the flexibility to adapt the methods to each specific context. Flexibility, however, needs to be accompanied by being cognizant of the complexities and impact that the adaptation process can have on the (functional) equivalence and thus comparability and validity of interpretations and fairness. Past studies such as Zhao & Solano-Flores (2020) illustrate how flexibility can be incorporated.

We have provided some insights into the strengths and potential challenges of each approach to test development, translation and adaptation. So far, the literature on international comparisons with PAs of CT has provided examples of studies following the formal condition design (Design I), the local condition design aiming at comparability (Design III-A), and the local condition design not aiming at comparability (Design III-B). However, there are no examples of studies following the concurrent development design (Design II). Additionally, the results of past studies, especially those that use different designs in the adaptation of a PA of CT, do not compare approaches. The comparison of such approaches, especially those aimed to achieve comparability in international assessments, would be a necessary step to inform and improve adaptation design choices. We hope that the framework presented here can stimulate future international studies of CT using PAs, and that it supports an informed choice of adaptation methods.

## REFERENCES

Addey, C., & Sellar, S. (2017). Why do countries participate in PISA? Understanding the role of international large-scale assessments in global education policy. In A. Verger, M. Novelli, H.K. Altinyelken (Eds.), *Global Education Policy and International Development: New Agendas, Issues and Policies* (pp. 97-117) London: Bloomsbury.

Aiken, M. (2019). An Updated Evaluation of Google Translate Accuracy. Studies in Linguistics and Literature. 3. p253. 10.22158/sll.v3n3p253.

Aiken, M., & Balan, S (2011). An Analysis of Google Translate Accuracy. *Translation Journal.* Retrieved June 20, 2021, from https://translationjournal.net/journal/56google.htm

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Basterra, M., Trumbull, E., & Solano-Flores, G. (2011). Cultural validity in assessment: Addressing linguistic and cultural diversity. NY: Routledge.

Bennet, R. (2020). Interpreting test-score comparisons. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 205–225). Washington, DC: National Academy of Education. Retrieved from https://naeducation.org/wp-content/uploads/2020/05/Comparability-of-Large-Scale-Educational-Assessments_final.pdf

Berman, A. I., Haertel, E. H., & Pellegrino, J. W. (2020). Introduction: Framing the issues. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 205–225). Washington, DC: National Academy of Education. Retrieved from https://naeducation.org/wp-content/uploads/2020/05/Comparability-of-Large-Scale-Educational-Assessments_final.pdf

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

Braun, M. (2006). Funktionale Äquivalenz in interkulturell vergleichenden Umfragen. Mythos und Realität [Functional equivalence in comparative intercultural surveys: myth and reality.] Mannheim: ZUMA.

Brown, G. (2016). Handbook of Human and Social Conditions in Assessment. In G. Brown, & L. Harris (Eds.) *Handbook of Human and Social Conditions in Assessment.* New York: Routledge. https://doi.org/10.4324/9781315749136

Byrne, B.M., Shavelson, R.J., & Muthen, B.O. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin*, 105(3), 456-466.

DePascale, C., & Gong, B. (2020). Comparability of individual students' scores on the "same test. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 205–225). Washington, DC: National Academy of Education. Retrieved from https://naeducation.org/wp-content/uploads/2020/05/Comparability-of-Large-Scale-Educational-Assessments_final.pdf

Ercikan, K., & Solano-Flores, G. (2016). Section Discussion: Assessment and Sociocultural Context: A Bidirectional Relationship. (pp. 490–505) In G. Brown, & L. Harris (Eds.) *Handbook of Human and Social Conditions in Assessment.* New York: Routledge. https://doi.org/10.4324/9781315749136

Ercikan, K., & Por, H.H. (2020). Comparability in multilingual and multicultural assessment contexts. In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 205–225). Washington, DC: National Academy of Education. Retrieved from https://naeducation.org/wp-content/uploads/2020/05/Comparability-of-Large-Scale-Educational-Assessments_final.pdf

Hambleton, R. K. (2002).Adapting achievement tests into multiple languages for international assessments. In A. C. Porter & A. Gamoran (Eds.), *Methodological advances in cross-national surveys of educational achievement* (pp. 58-79). Washington, DC: National Academy Press.

Hambleton, R. K. (2005) Issues, Designs, and Technical Guidelines for Adapting Tests Into Multiple Languages and Cultures. In R.K. Hambleton, P.F. Merenda, C.D. Spielberger (Eds.) Adapting Educational and Psychological Tests for Cross-Cultural Assessment. Psychology Press.

Hambleton, R. K., & Zenisky, A. L. (2010). Translating and adapting tests for cross-cultural assessments. Cross-Cultural Research Methods in Psychology In D. Matsumoto & F. Van de Vijver (Eds.), *Cross-Cultural Research Methods in Psychology* (pp. 46-70). Cambridge: Cambridge University Press. doi:10.1017/CBO9780511779381.004

Hyytinen, H., & Toom, A. (2019). Developing a performance assessment task in the Finnish higher education context: Conceptual and empirical insights. *British Journal of Educational Psychology, 89*(3), 551–563. https://doi.org/10.1111/bjep.12283

Holtsch, D., Rohr-Mentele, S., Wenger, E., Eberle, F., & Shavelson, R. J. (2016). Challenges of a cross-national computer-based test adaptation. *Empirical research in vocational education and training, 8(18)*, 1–32. https://doi.org/10.1186/s40461-016-0043-y

International Test Comission. (2017). *The ITC Guidelines for Translating and Adapting Tests (Second edition)*. https://doi.org/10.1111/j.1464-0597.1975.tb00322.x

Keng, L., & Marion, S. (2020). Comparability of aggregated group scores on the "same test". In A. I. Berman, E. H. Haertel, & J. W. Pellegrino (Eds.), *Comparability of large-scale educational assessments: Issues and recommendations* (pp. 205–225). Washington, DC: National Academy of Education. Retrieved from https://naeducation.org/wp-content/uploads/2020/05/Comparability-of-Large-Scale-Educational-Assessments_final.pdf

Liu, O. L., Frankel, L., & Roohr, K. C. (2014). Assessing critical thinking in higher education: current state and directions for next-generation assessments. *ETS Res. Rep. Ser. 1,* 1–23. https://doi.org/10.1002/ets2.12009

McClelland, D. C. (1973). Testing for competence rather than for "intelligence." *American Psychologist, 28(1)*, 1–14. https://doi.org/10.1037/h0034092

Mihailidis, P., & Thevenin, B. (2013). Media Literacy as a Core Competency for Engaged Citizenship in Participatory Democracy. *American Behavioral Scientist, 57(11),* 1611–1622. https://doi.org/10.1177/0002764213489015

Mislevy, R. J. (2018). *Sociocognitive foundations of educational measurement*. New York: Routledge. https://doi.org/10.4324/9781315871691

O'Leary, M., Reynolds, K., Guangming, L.; Ou, L.L., Belton, S., O'Reilly, N., & McKenna, J., (2020). Assessing Critical Thinking in Higher Education: Validity Evidence for the Use of the HEIghten™ Critical Thinking Test in Ireland. Journal of Higher Education Theory & Practice. Vol. 20 Issue 12, p115-130

Oser, F. K., & Biedermann, H. (2020). A three-level model for critical thinking: critical alertness, critical reflection, and critical analysis. In O. Zlatkin-Troitschanskaia (Eds.) *Frontiers and Advances in Positive Learning in the Age of Information (PLATO)*. (pp. 89–106). Springer. https://doi.org/10.1007/978-3-030-26578-6_7

Perie, M. (2020). Comparability across different assessment systems. In A.I. Berman, E.H. Haertel, J.W. Pellegrino (Eds.), *Comparability of Large-Scale Educational Assessments: Issues and Recommendations*. National Academy of Education. Retrieved from https://naeducation.org/wp-content/uploads/2020/05/Comparability-of-Large-Scale-Educational-Assessments_final.pdf

Rogers, W. T., Gierl, M. J., Tardif, C., Lin, J., & Rinaldi, C. (2003). Differential Validity and Utility of Successive and Simultaneous Approaches to the Development of Equivalent Achievement Tests in French and English. *Alberta Journal of Educational Research, 49(3*), 290–304.

Rogers, W. T., Lin, J., & Rinaldi, C. (2011). Validity of the simultaneous approach to the development of equivalent achievement tests in English and French. *Applied Measurement in Education, 24*, 39-70.

Schendel, R., & Tolmie, A. (2017). Beyond translation: adapting a performance-task-based assessment of critical thinking ability for use in Rwanda. Assessment and Evaluation in Higher Education, 42(5), 673–689. https://doi.org/10.1080/02602938.2016.1177484.

Shavelson, R.J. (2010). *Measuring college learning responsibly: Accountability in a new era.* Stanford, CA: Stanford University Press.

Shavelson, R.J., Solano-Flores, G., & Kurpius, A. (2010). *GS.4 Conceptual Framework. AHELO Module A: Adaptation and Translation of Performance Tasks.* Council for Aid to Education.

Shavelson, R.J., Zlatkin-Troitschanskaia, O., Mariño, J. (2018). International Performance Assessment of Learning in Higher Education (iPAL): Research and Development. In: O. Zlatkin-Troitschanskaia, M. Toepper, H. Pant, C. Lautenbach, C. Kuhn (Eds.) *Assessment of Learning Outcomes in Higher Education. Methodology of Educational Measurement and Assessment.* Springer, Cham. https://doi.org/10.1007/978-3-319-74338-7_10

Shavelson, R.J., Zlatkin-Troitschanskaia, O., Beck, K., Schmidt, S., Mariño, J. (2019). Assessment of University Students' Critical Thinking: Next Generation Performance Assessment, *International Journal of Testing 19:4,* 337-362. https://doi.org/10.1080/15305058.2018.1543309

Siegel, H. (2010). On Thinking Skills. In C. Winch (Eds). *Teaching thinking skills* (2nd ed.). New York, NY : Continuum International Pub. Group.

Sierci, S., & O'Riordan, M. (2020). Comparability when assessing individuals with disabilities. In A.I. Berman, E.H. Haertel, J.W. Pellegrino (Eds.), *Comparability of Large-Scale Educational Assessments: Issues and Recommendations*. National Academy of Education. Retrieved from https://naeducation.org/wp-content/uploads/2020/05/Comparability-of-Large-Scale-Educational-Assessments_final.pdf

Solano-Flores, G. (2011). Assessing the Cultural Validity of Assessment Practices: An Introduction. In M.R. Basterra, E. Trumbull, G. Solano-Flores. *Cultural validity in assessment: Addressing linguistic and cultural diversity.* NY: Routledge. https://doi.org/10.4324/9780203850954

Solano-Flores, G. (2012).*Translation Accommodations Framework for Testing English Language Learners in Mathematics.* Smarter Balanced Assessment Consortium, September 18, 2012. Retrieved from https://portal.smarterbalanced.org/library/en/item-accessibility-and-language-variation-conceptual-framework.pdf

Solano-Flores, G. (2019a). Examining Cultural Responsiveness in Large-Scale Assessment: The Matrix of Evidence for Validity Argumentation. *Frontiers in Education, 4(June)*, 1–9. https://doi.org/10.3389/feduc.2019.00043 Retrieved from https://www.frontiersin.org/articles/10.3389/feduc.2019.00043/full

Solano-Flores, G. (2019b). The participation of Latin American Countries in International Assessments: Assessment Capacity, Validity, and Fairness. In, L. E. Suter, E. Smith & B. D. Denman, B. D.T (Eds.), *Sage Handbook on Comparative Studies in Education: Practices and Experiences in student schooling and learning* (pp. 139-161). Thousand Oaks, CA: Sage.

Solano-Flores, G., Javanovic, J., Shavelson, R. J., & Bachman, M. (1999). On the development and evaluation of a shell for generating science performance assessments. *International Journal of Science Education, 21(3),* 293–315. https://doi.org/10.1080/095006999290714

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent Development of Dual Language Assessments: An Alternative to Translating Tests for Linguistic Minorities. *International Journal of Testing, 2(2),* 107–129. https://doi.org/10.1207/s15327574ijt0202_2

Solano-Flores, G., Backhoff, E., & Contreras-Niño, L. Á. (2009). Theory of Test Translation Error. *International Journal of Testing, 9(2),* 78–91. https://doi.org/10.1080/15305050902880835

Solano-Flores, G., Chía, M., Shavelson, R.J., & Kurpius, A. (2010). *GS.36. Translation Guide. AHELO Module A.* Council for Aid to Education.

Solano-Flores, G., Shade, C., & Chrzanowski, A. (2014). *Item accessibility and language variation conceptual framework. Submitted to the Smarter Balanced Assessment Consortium. October 10. Retrieved June 11, 2021 from https://portal.smarterbalanced.org/library/en/item-accessibility-and-language-variation-conceptual-framework.pdf*

Survey Research Center, Institute for Social Research, University of Michigan (2016). *Guidelines for Best Practice in Cross-Cultural Surveys.* http://ccsg.isr.umich.edu/images/PDFs/CCSG_Full_Guidelines_2016_Version.pdf

Stecher, B. M., Klein, S. P., Solano-Flores, G., McCaffrey, D., Robyn, A., Shavelson, R. J., & Haertel, E. (2000). The Effects of Content, Format, and Inquiry Level on Science Performance Assessment Scores. Applied Measurement in Education, 13(2), 139–160.

Stoltz, B. (n.d.). Google Translate continues to improve, but how accurate is it? Retrieved June 20, 2021, from: https://burgtranslations.com/google-translate-continues-to-improve-but-how-accurate-is-it/

Taira, B. R., Kreger, V., Orue, A., & Diamond, L. C. (2021). A Pragmatic Assessment of Google Translate for Emergency Department Instructions. *Journal of General Internal Medicine.* https://doi.org/10.1007/s11606-021-06666-z

Tremblay, K., Lalancette, D., & Roseveare, D. (2012). *Assessment of higher education learning outcomes (AHELO): Rationale, challenges and initial insights from the feasibility study.* OECD (Vol. 1). https://doi.org/10.1007/978-94-6091-867-4 http://www.oecd.org/education/skills-beyond-school/AHELOFSReportVolume1.pdf

van de Vijver, F. J. R., & Poortinga, Y. H. (2004). Conceptual and methodological issues in adapting tests. *Adapting Educational and Psychological Tests for Cross-Cultural Assessment*, 39–63. https://doi.org/10.4324/9781410611758

van de Vijver, F. J. R., & Poortinga, Y. H. (2016). On item pools, swimming pools, birds with webbed feet, and the professionalization of multilingual assessment. In C. S. Wells & M. Faulkner-Bond (Eds.), *Educational measurement: From foundations to future (*pp. 273–290). The Guilford Press.

van de Vijver, F. J. R. (2016). Assessment in education in multicultural populations. In G. T. L. Brown and L. Harris (Eds) *Handbook of Human and Social Conditions of Assessment.* (pp. 436–453). New York, NY: Routledge

Zhao, X., & Solano-Flores, G. (2020). Testing across languages in international comparisons: cultural adaptation of consensus-based test translation review procedures, *Journal of Multilingual and Multicultural Development.* https://doi.org/10.1080/01434632.2020.1852242

Zahner, D., & Ciolfi, A. (2018). International Comparison of a Performance-Based Assessment in Higher Education. In: O. Zlatkin-Troitschanskaia, M. Toepper, H.A. Pant, C. Lautenbach and C. Kuhn (Eds.), *Assessment of Learning Outcomes in Higher Education Cross-National Comparisons and Perspectives.* Springer. https://doi.org/10.1007/978-3-319-74338-7_11.

Zlatkin-Troitschanskaia, O., Toepper, M., Molerov, D., Buske, R., Brückner, S., Pant, H., Hofmann, S., & Hansen-Schirra, S. (2018). Adapting and Validating the Collegiate Learning Assessment to Measure Generic Academic Skills of Students in Germany: Implications for

International Assessment Studies in Higher Education. In: O. Zlatkin-Troitschanskaia, M. Toepper, H.A. Pant, C. Lautenbach and C. Kuhn (Eds.), *Assessment of Learning Outcomes in Higher Education Cross-National Comparisons and Perspectives*. Springer. https://doi.org/10.1007/978-3-319-74338-7_12.

## ABOUT THE AUTHORS

***Natalia Ronderos***

Natalia Ronderos is a doctoral student of the joint doctoral program in Education at the University of Zürich and the St. Gallen University of Teacher Education (PHSG). Her research is on international performance assessments of critical thinking and is situated on teacher education. Her experience includes test development work and leading a study on test preparation effects on SABER 11 (high-school exit exam used for college admission) at ICFES, the national assessment system of Colombia. Additionally, she has several years of involvement in higher education institutions on teaching and its evaluation. More recently, she was Deputy Director of the Center for Evaluation of Education of Universidad de Los Andes, where she was co-investigator in adapting a Finnish performance assessment of critical thinking to the Colombian context within the iPAL framework. For several years Natalia was Director of Teaching and Instruction at Universidad Jorge Tadeo Lozano, leading several initiatives to promote the scholarship of teaching and learning. She received her bachelor's degrees in sociology and economics from Universidad de Los Andes and Universidad Nacional. She holds a master's degree in Education from Stanford University.

**Contact information:** University of Zürich, natalia.ronderosbarreto@uzh.ch

***Richard J. Shavelson***

Richard J. Shavelson is Professor of Education and Psychology, Dean of the Graduate School of Education and Senior Fellow in the Woods Environmental Institute (Emeritus) at Stanford University. He was president of AERA; a fellow of American Association for the Advancement of Science, AERA, American Psychological Association, and the American Psychological Society; a Humboldt Fellow; and member of National Academy of Education and International Academy of Education. His work focuses on performance assessment of undergraduates' learning. His publications include Statistical Reasoning for the Behavioral Sciences, Generalizability Theory: A Primer, Scientific Research in Education; Assessing College Learning Responsibly: Accountability in a New Era.

**Contact information:** Stanford University, 650-868-1811, richs@stanford.edu

***Doreen Holtsch***

Doreen Holtsch has been Director of the Institute of Research on Teaching Profession and on Development of Competencies at the St.Gallen University of Teacher Education, Switzerland, since 2018. Doreen Holtsch completed her studies in business and economics education at the Humboldt University in Berlin and the subsequent doctoral studies at the University of Rostock. She gained her postdoctoral qualification in 2018 at the University of Zurich, where she has been the operational manager of the Leading House for Vocational Education Research "Learning and Instruction for Commercial Apprentices" (LINCA). She has carried various teaching and research activities at universities in Switzerland and Germany dealing with teaching and learning processes at (vocational) schools and companies. Amongst them she was responsible for a cross-national adaptation study of a computer-based test.

**Contact information:** St.Gallen University of Teacher Education, Notkerstrasse 27, 9000 St. Gallen, +41 71 243 96 30, doreen.holtsch@phsg.ch

***Olga Zlatkin-Troitschanskaia***
Professor Olga Zlatkin-Troitschanskaia has been Chair of Business and Economics Education at Johannes Gutenberg University Mainz (JGU), Germany, since 2006. She earned her doctoral degree from Humboldt University of Berlin in 2004 and her postdoctoral qualification in 2006. She has published widely on empirical educational research in vocational and higher education. She has directed numerous externally funded national and international research projects and has been coordinating the national research program 'Modeling and Measuring Competencies in Higher Education (KoKoHs)' since 2011; she also co-implemented the international collaborative research project 'Performance Assessment of Learning' (iPAL). Her research has earned various awards and honors. She is a member of many national and international research academies as well as advisory and editorial boards, and serves as an expert consultant to ministries, foundations, and academic journals.
**Contact information:** Johannes Gutenberg University, Jakob-Welder-Weg 9, 55128 Mainz Germany, +49 6131 39-22009, troitschanskaia@uni-mainz.de

***Guillermo Solano-Flores***
Guillermo Solano-Flores is Professor of Education at the Stanford University Graduate School of Education. He specializes in educational assessment and the linguistic and cultural issues that are relevant to both international test comparisons and the testing of cultural and linguistic minorities. His research is based on the use of multidisciplinary approaches that use psychometrics, sociolinguistics, semiotics, and cognitive science in combination. He is the author of the theory of test translation error, which addresses testing across cultures and languages. Also, he has investigated the use of generalizability theory—a psychometric theory of measurement error—in the testing of English language learners and indigenous populations. He has advised Latin American countries on the development of national assessment systems. Also, he has been the advisor to countries in Latin America, Asia, Europe, Middle East, and Northern Africa on the adaptation and translation of performance tasks into multiple languages, and has proposed the concept of national assessment capacity as critical for countries to benefit from their participation in international test comparisons.
**Contact information:** Stanford University, 485 Lasuen Mall, Stanford, CA 94305-3096, 650 -723-2109, gsolanof@stanford.edu