Research Report

The Validity of Inferences From Locally Developed Assessments Administered Globally

ETS RR–18-35

María Elena Oliveri
René Lawless

December 2018

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

The Validity of Inferences From Locally Developed Assessments Administered Globally

María Elena Oliveri & René Lawless

Educational Testing Service, Princeton, NJ

In this paper, we first examine the challenges of score comparability associated with the use of assessments that are exported. By exported assessments, we mean assessments that are developed for domestic use and are then administered in other countries in either the same or a different language. Second, we provide suggestions to better support their valid and fair use. We illustrate these issues in the context of higher education assessments that are designed to serve different purposes—inform admissions decisions and assess student learning outcomes within one country (e.g., the United States) that are later used in other countries. In higher education, the use of exported assessments is on the rise due to increases in globalization, student mobility, and cross-national comparisons of student achievement. An increase in the use of exported assessments leads to more diverse test-taker populations and requires special attention due to possible sources of construct-irrelevant variance, which may threaten the score-based inferences made for various populations. Irrelevant sources of variance may emerge due to differences in opportunity to learn, curricular exposure, and lack of familiarity with the cultural references used in the assessments that are exported.

Keywords Exported assessments; repurposed assessments; international assessments; validity; construct-irrelevant variance

doi:10.1002/ets2.12221

Increased globalization and student mobility are contributing to an increased use of *exported assessments*. Exported assessments are originally developed for domestic use in one country and then administered in other countries to inform a similar use (e.g., postsecondary admissions or selection decisions). They may be exported to countries in the same or a different language. To support fair and valid score-based inferences, their use in new countries requires analyses of curricular, content, linguistic, and sometimes cultural relevance.

Wendler and Powers (2009) highlighted the fact that validity threats may arise when repurposing assessments (i.e., using assessments with new populations, which may differ from the original populations for which the test was initially developed). Ercikan and Lyons-Thomas (2013), Hambleton, Merenda, and Spielberger (2005), and Oliveri, Ercikan, and Simon (2015) have argued that the use of assessments with multiple populations requires the careful consideration of cultural, linguistic, sociocultural, ecological, and curricular differences across the populations to which the assessment will be administered. Although previous researchers have pointed to the importance of evaluating various issues when administering assessments with multiple populations, further research is needed to let us know what actions could help improve score-based inferences for the multiple populations and when such steps could be taken (e.g., during test development, use, or interpretation of score processes). Moreover, although standards have been developed to guide such inferences, they are primarily useful to practitioners who are already experts in test adaptation or related practices because they are not often written in accessible language and often lack examples to render the information more practical and provide additional guidance regarding how to apply standards (e.g., American Educational Research Association [AERA], American Psychological Association [APA], and the National Council on Measurement in Education [NCME] *Standards for Educational and Psychological Testing* (2014); ITC, 2010) to the development and use of exported (repurposed) assessments.

In this paper, we outline challenges that may arise from exporting an assessment. Examples of such challenges are threats to score comparability, fairness, and validity of score-based interpretations. Also, we provide practical suggestions to help enhance test development practices when the assessment is exported. The suggestions go beyond traditional practices of item translation or adaptation as the primary step taken to establish score comparability. Instead, we provide a systematic view of best practice approaches along key stages of test development, score use, and interpretation in the analysis of score comparability issues in exported assessments. We suggest this need exists because current guidelines and

Corresponding author: María Elena Oliveri, E-mail: moliveri@ets.org

frameworks remain fairly general in relation to the steps needed to guide the principled development, use, and interpretation of exported assessments.

To illustrate the relevant steps needed to help support valid score-based interpretations of exported assessments, we use a postsecondary education context in relation to two assessment use-case scenarios. In the first scenario, we examined tests used for higher education admissions. Such tests are typically high stakes and result in individual scores. Examples include the ACT, the *SAT*®, the *Graduate Record Examinations*® (*GRE*®) General Test, and the Law School Admission Test (LSAT). Second, we examined student learning outcomes (SLO) assessments, which are often low stakes for the individual test takers and in which the results are analyzed at an aggregated level (group or institutional) for curricular evaluation and overall course mastery; an example is the *ETS*® Major Field Tests. We frame this research in the postsecondary education context because the use of exported assessments is increasing due to growing numbers of international applicants applying to higher education institutions, according to Educational Testing Service (ETS, 2014c) and because higher numbers of postsecondary institutions want to use a *common measure* to inform decisions about admissions and program effectiveness and to benchmark their performance against other programs in other countries (ETS, 2014a).

Data Sources (Assessment Use Scenarios)

The steps we outline in this paper are in accordance with the International Test Commission (ITC, 2018) *ITC Guidelines for Large-Scale Assessments of Linguistically Diverse Populations*, which indicate that fairness and validity need to be examined throughout the assessment process (from test conceptualization and design through the reporting and interpretations of scores). The proposed approach is also in agreement with the AERA, APA, and NCME (2014), which recommends the use of "test design, development, administration, and scoring procedures that minimize barriers to valid score interpretations for the widest possible range of individuals and relevant subgroups" (p. 63). These documents were essential references in our investigation. Other sources included test development and test administration guidelines for the assessment of English language learners (ETS, 2014b; Pitoniak et al., 2009; Young, So, & Ockey, 2013), for test use (ITC, 2000, 2001, 2010, 2018), and the test exportation framework (Oliveri, Lawless, & Young, 2015). These publications helped us identify the kinds of considerations that are relevant to the evaluation of exported assessments.

In particular, two ITC (2010) guidelines illustrate the centrality of considering linguistic and cultural factors when assessing diverse populations. These guidelines (C1 and D1) state that the "Effects of cultural differences which are not relevant or important to the main purposes of the study should be minimized to the extent possible." The second guideline states that "Test developers/publishers should insure that the adaptation process takes full account of linguistic and cultural differences among the populations for whom adapted versions of the test or instrument are intended" (p. 2). These guidelines suggest that the linguistic and cultural ways of contextualizing test items can interact with a test taker's understanding of, and access to, the test content.

Another source of information came from periodic meetings (bimonthly meetings over a 12-month period) with a team of assessment experts. The team consisted of researchers, psychometricians, and assessment development specialists. Team members had expertise in procedures and processes related to our two assessment contexts and provided insights regarding best practices in the international use of assessments as well as practical suggestions to help enhance the validity of the score-based inferences derived from exported assessments.

Considerations for Enhancing the Validity of Exported Assessments

In this section, we discuss considerations and suggestions for enhancing the validity of score-based inferences for exported assessments used to help assess SLO and inform admissions decisions. The presented examples are not meant to be exhaustive of all possible sources of construct-irrelevant variance that may lead to alternative score-based interpretations, nor do we intend for our suggestions to be prescriptive. As mentioned, the degree to which our suggestions can be implemented depend on a test developer's resources in terms of financial, staffing, and other considerations that might make it relevant or useful to adopt particular suggestions and not others. Our suggestions are informed by (a) our literature review, (b) our evaluation of items and ancillary materials from exported assessments, and (c) collaborations with a team of content experts. The goal is to enhance assessments' validity and fairness in support of principled exported use.

We organize our suggestions along six components related to test development, use, and score-based interpretations. These components are *domain definition, evaluation, generalization, explanation, extrapolation,* and *utilization*. We

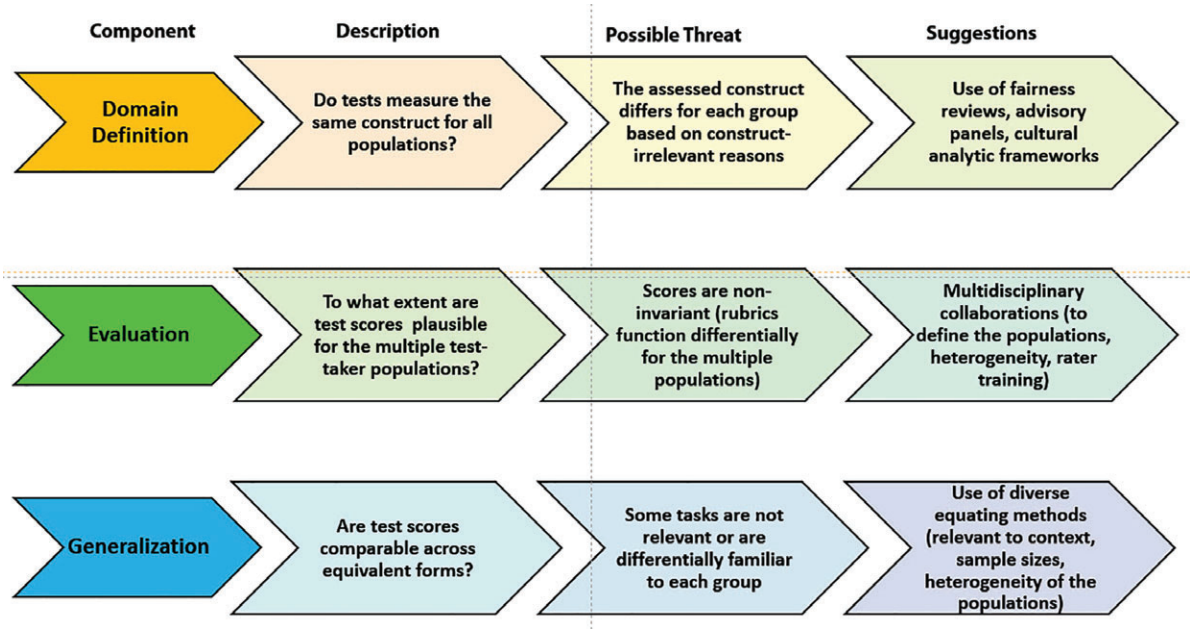

Figure 1 Consideration of possible threats arising in the domain definition, evaluation, and generalization components of developing and using assessments that are exported.

evaluate the possible threats and provide recommendations for each one in light of Kane's (2013) argument-based approach to validation. Kane's approach has previously been applied to language-proficiency assessments such as the *TOEFL*® test according to Chapelle (2008); Chapelle, Enright, and Jamieson (2010); and Xi (2010), but which had not yet been evaluated in the case of using exported assessments as described in this paper.

Domain Analysis, Evaluation, and Generalization Components

Figure 1 summarizes the first three components (domain definition, evaluation, and generalization). It provides a brief description of the components, their roles in the argument, examples of threats, and suggestions for dealing with the threats. These three components surround considerations related to test conceptualization and design.

Component 1: Domain Definition

Threats that may be identified in the domain definition component may include whether (a) the test measures the intended construct for all populations, (b) the test takers have had an opportunity to learn the assessed curriculum covered in the test, and (c) the items or tasks contain possible sources of construct-irrelevant variance that may lead to differential score-based inferences for the different populations.

A goal of the domain definition is to design an assessment that measures the targeted knowledge, skills, and abilities (KSAs) for all populations. In admissions, the assessments should measure constructs identified as relevant to the purposes of the assessment (e.g., assessing general competence related to written or oral communication, mathematics, or analytical reasoning). Key considerations for meaningful interpretation of exported assessments' scores include whether or not the assessed constructs are universal, similarly useful, or both to populations in the countries to which the assessment will be exported. If the test will be used to make comparisons across countries, such as evaluating program effectiveness or informing curricular reforms for separate countries, efforts are needed to help ensure that the test covers similar curricula across countries. If the test is developed from the ground up with a perspective of use across multiple countries, representatives (curricular experts) from each country can help inform the test blueprints so that the test stays relevant and current for institutions outside of the test development country. A curricular survey may be developed and reviewed by a test development committee that includes both national and international representatives to help ensure more generalizable curriculum content and minimize the risk that the test is not solely reflecting the test developer's culture.

For instance, a U.S.-developed test used for admission into a U.S. law school may focus on skills that are specifically relevant to legal practice in the United States but are less relevant to legal practice in other countries. Such uses may be

acceptable for admissions into a U.S. law school but not for admission into a law school in a different country, whereas other (more general) aspects of the test may be deemed suitable for both countries. Along the same lines, Ling (2013) provided an example with respect to an SLO assessment that tests business skills: "the content covered or focused on in non-U.S.-based business programs may slightly differ from that in the United States, mainly because the business context in these countries and regions differs from the U.S. context to varying degrees" (p. 9). In such instances, curricular experts in the new country may want to identify which components are applicable to their country and which ones are less so. Comparisons across countries should be informed by the items that are comparable and content that students in both countries are familiar with and have had exposure to. Arguably, tests that have little (e.g., 50% or less) overlap may not be suitable for exportation unless the test takers who will be taking the assessment are provided with review material that explicitly targets those aspects of the curriculum that are different across countries.

A second goal of the domain analysis is to identify potential sources of construct-irrelevant variance that could lead to differential score-based inferences for the different populations. This process can be informed by meeting with experts (e.g., assessment developers, content experts, or both) who have familiarity with the language and culture of the new, targeted population(s) and can review test items for fairness as well as point to particular task types that may be differentially familiar to the new populations. Their evaluation of items might entail looking for different item features that could introduce unnecessary sources of complexity such as the use of idiomatic expressions or units of measurement that may be differentially familiar to some cultures as compared to others. An example would be an item that asks students to openly demonstrate ideas, behaviors, and possibly disagreement. For example, Cheung (2004) and Cheung et al. (1996) suggested that Chinese test takers may be less inclined than American test takers to openly and directly express behaviors such as anger or disagreement. This issue is relevant to the assessment of written competency and the inclusion of items that require individuals to agree or disagree with a government law such as allowing online voting. Although in some countries test takers may feel free to openly disagree with the government on a law, such open disagreement may be unacceptable in other countries. The inclusion of such items may introduce construct-irrelevant differences across test-taker populations. Test developers may want to consider differential cultural tendencies when adapting items for multiple populations for meaningful score-based inferences (Oliveri, Lawless, & Mislevy, under review).

Table 1 provides an overview of possible sources of construct-irrelevant variance that could threaten valid score interpretations for different populations. Examples of such threats include differences in familiarity with an assessed curriculum (Dwyer, Gallagher, Levin, & Morley, 2003), item types used in a test (Huff, 2000), different degrees of motivation (Maddox, 2015), and the comprehension of geocultural references and terms used in the test items or stimuli (Oliveri et al., 2018). There could also be threats introduced by differences in item type familiarity, the presence of idioms, and geocultural terms.

We now propose ways in which experts can help inform the fairness review process in more detail. For instance, curricular, cultural, and linguistic experts can help identify possible sources of construct-irrelevant variance for the major test-taker populations prior to exporting an assessment (Roth, Oliveri, Sandilands, Lyons-Thomas, & Ercikan, 2013). The group of reviewers may contain subject matter experts (e.g., appropriate faculty members, administrators, or both) to undertake a content review to determine whether the content and coverage of the tests are consistent with the content covered in the various curricula and aligned with the expectations of students majoring in the particular field(s) to which the test applies.

We suggest that these reviewers should be familiar with (a) the original and targeted test populations, (b) specific phrases or idioms used in the language in which the test was developed, and (c) the linguistic features of the items that could be problematic. These reviewers would provide an inherent point of view about aspects of items that may be unfamiliar to international test takers and therefore possibly construct irrelevant. These fairness reviewers may also examine (post hoc) those items in which differential performances between domestic and international populations have been noted, those for which the performances are very similar, and perhaps those items that are removed from the item pool because of high levels of differential item functioning (DIF).[1]

Table 1 provides an example of the features of items that can be analyzed. Additional examples relevant to the fair assessment of multiple populations are provided in guidelines such as the *ETS Standards for Quality and Fairness* (ETS, 2014b), which suggests avoiding the use of cognitive sources of construct-irrelevant variance by eliminating unnecessarily difficult language from an assessment. Examples are the use of topics that may include the use of regionalisms, references to religion(s), and references to a particular country's national culture. Examples from these sources can be used to train

Table 1 Sources of Construct-Irrelevant Differences That May Lead to Performance Differences Between Populations

Examples of possible sources of construct-irrelevant variance	Description	Examples
Item language—differential use of idioms	A phrase unique to a given culture, the meaning of which is figurative, not literal	• "pay lip service to" • "struck by the fact" • "flush with"
Item language—differential use of proper names	The names of people, actors, or other figures that are associated with a particular culture (Oliveri, Lawless, Robin, & Bridgeman, 2018).	• Mary Louise Parker • Navajo • Meriwether Lewis
Cultural differences (differential understanding of, and familiarity with, geocultural references)	Names of cities, towns, geographical locations, or processes that are associated with a particular type of government (Oliveri et al., 2018; Young et al., 2014)	• Mississippi • A particular county • Peculiarities of national legal systems
Test materials	Differential familiarity with item types (Oliveri et al., 2018)	• Quantitative comparisons • Analogies • Sentence equivalence
Test-taking behaviors	Different approaches to acquiescence, unanswered items, tendency toward guessing, and cognizance of social desirability in the test-taking environment (Maddox, 2015)	• The different conceptualizations of "cheating" across cultures
Levels of motivation	The degree of familiarity with text types (reading texts or stimuli) might differentially impact test takers' motivation as lower degrees of familiarity may pose additional challenges to international test takers (Oliveri et al., 2017).	• The use of unfamiliar contexts in reading stimuli
Scoring of constructed-response items	Cultural differences related to writing style, differential word use, and succinctness (Oliveri, Lawless, & Young, 2015). Prose that is succinct versus prose that is very descriptive may be culturally dependent in terms of how communication is structured.	• Very long explanatory essays full of colorful language versus compact essays that focus on the main points only

reviewers for fairness and may include examples of items that illustrate potentially problematic language, such as items that contain geocentric language (Oliveri et al., 2018). Reviewers can suggest minor adaptations—for instance, to ensure that units of measurement (metric vs. English) are familiar to all test takers to enable them to demonstrate their full range of construct-relevant KSAs.

Component 2: Evaluation

The analyses in the evaluation component involve examining the degree to which the observed scores derived from a test are plausible and appropriate for their proposed use and reflect the targeted construct for all populations. Various steps can be followed based on the kinds of methods used for item analysis across populations. For instance, analyses should be conducted to examine whether the rubrics used for scoring test responses are appropriate; such analyses may also involve examining the degree to which constructed-response (CR) items, including scoring rubrics, are free of possible sources of construct-irrelevant variance.

The evaluation of the scoring rubrics for CR items involves asking whether the rubrics are designed to capture the construct of interest and allow valid score-based inferences for all test-taker populations. Several studies that evaluate the validity of CR items may point to possible threats underlying differential item-response patterns across groups (clustered

by language or by country). These sources may include differential degrees of topic familiarity, differences in the wording of prompts, scoring mode, type of variant, and the rating approaches of CR items (essays). For example, results of a study conducted by Shi (2001) revealed differences across raters in the scoring of English CR items answered by Chinese university students. This appeared to be related to whether the raters were native English, English as a Foreign Language (EFL) teachers, or nonnative EFL teachers. The two groups of teachers neither gave similar scores to the same writing task nor used the same criteria in their scoring of the essays. More specifically, the native English EFL teachers focused on content and language, whereas the Chinese (nonnative EFL teachers) were more concerned with and gave more weight to the essay's organization and length. Along the same lines, Eckes (2008), Uysal (2008), and Zhan (1992) suggested that the background of raters influences the scores they give, including the degree of leniency or severity, weight given to language errors, and allowance for organization and rhetoric in writing patterns (e.g., the use of explicit transition cues or the test taker's willingness to take an unambiguous stance).

Differences across scoring mode (i.e., human vs. automated scoring) might also arise when using exported assessments of multiple populations. Bridgeman, Trapani, and Attali (2012) identified differences across language groups (e.g., Arabic and Chinese), as the latter group had higher mean score differences from automated scoring compared to human raters for different kinds of CR prompts. The score discrepancies may have been caused by differences in essay length or in writing styles between the language groups. To safeguard against the introduction of possible sources of construct-irrelevant variance in the scoring process, assessment programs that score CR items should consider various calibration processes to help ensure consistency and accuracy across raters. For example, raters may be provided with extensive instructions on essay scoring, receive extensive training or certification, and may view benchmark papers as exemplars during the training.

Sample size permitting, one may also consider conducting DIF analyses of the new populations. If DIF is found, one may want to analyze its sources (e.g., using cognitive interviews, which we will elaborate on later in the paper) to determine whether or not the DIF is construct relevant or irrelevant. Various challenges may arise when conducting DIF analyses to compare the responses of the original and new populations. One challenge is how to classify test takers into groups (e.g., by world region, by language group, by geographic proximity, or on all international test takers combined). Sinharay, Dorans, and Liang (2009) suggested conducting DIF analyses by grouping examinees by first-language proficiency, as a higher (or lower) number of test takers who report that English is not their first language may lead to different DIF results. Another approach is to group examinees by language groups. Oliveri and von Davier (2016) exemplified ways to group examinees by language groups, such as Indo-European versus non-Indo-European languages, or combining members of the same language family (e.g., grouping all Romance languages together).

Challenges in DIF analysis may also arise when sample sizes are small. Various methods of analyzing DIF with small sample sizes have been proposed. Examples of these methods include Bayesian estimation approaches (Zwick, Thayer, & Lewis, 1999, 2000), which build upon the Mantel–Haenszel (MH) method, or the delta plot and the standardization index methods (Dorans & Kulick, 1986; Muñiz, Hambleton, & Xing, 2001) as well as the random-groups equating method, which can function with samples as low as 50 to 400 test takers.

Another consideration is the analysis of within-group heterogeneity. Research in DIF analysis, including the use of latent class methodologies (von Davier, 2008), have been proposed to examine DIF with heterogeneous (e.g., language and cultural) groups (Oliveri, Ercikan, & Zumbo, 2013). Similarly, extensions of the Zhang, Dorans, and Matthews-López (2005) DIF dissection approach, which involves the creation of more precisely defined groups by creating finer groupings within larger (total) groups (e.g., gender crossed by ethnicity) have been applied to language groups (Ercikan & Oliveri, 2013). The authors contended that DIF dissection leads to more accurate DIF results for investigations of item-response patterns for heterogeneous groups.

Besides DIF analyses, field tests are also proposed as a way to identify possible threats to score comparability with international populations. Their importance is mentioned across various standards and guidelines. Pitoniak et al. (2009) suggested conducting small-scale pilot tests with a sample of test takers who are representative of the target population. These studies are important because the resulting data can indicate whether the items are understood in the way they were intended or if revisions to the test items need to be made due to an increased diversity in the demographics of the population. They can also serve to alert content and fairness reviewers to items that contain geocentric wording or content or possibly unclear instructions for the new population.

Timing requirements may also need to be examined for possible *speededness*. Speededness is a possible source of construct-irrelevant variance that indicates that test takers are not able to fully demonstrate their KSAs due to time

constraints (Spielberger, Moscoso, & Brunner, 2005). Field tests can be conducted in collaboration with test developers, psychometricians, and researchers for help in interpreting score differences or DIF results that may become apparent. Early DIF detection can help ensure that the items are reviewed by experts and, if necessary, are appropriately revised for subsequent operational administrations of the assessment.

If DIF analysis suggests that there are possible timing differences, speededness, or other issues between the original and new populations in their item responses, cognitive reviews of items can be conducted. For instance, such interviews may yield useful information with regard to possible differences in the way test questions affect test takers' thinking processes. *Are the items understood by the test takers as intended? Do the items elicit different responses because of the introduction of construct-relevant* versus *construct-irrelevant contexts?* Cognitive interviews and think-aloud protocols are useful approaches for empirically examining test takers' cognitive processes, for understanding the measured constructs, and for gathering assessment validity evidence (Baxter & Glaser, 1998; Ercikan et al., 2010).

Given their one-on-one approach, cognitive interviews cannot accommodate large sample sizes but can be useful in gathering qualitative data about how different aspects of test items affect test takers' thinking processes and test performance; they also lend themselves well to comparing variations in test-taker thought processes across language groups. However, note that one limitation of cognitive interviews is whether or not the information captured from a subset of test takers will generalize to all test takers. They can be useful in identifying potential sources of DIF, any hidden assumptions, or alternative plausible interpretations of responses to the items that were already administered in tests. The information obtained in these interviews can shed light on differences in thought patterns between populations and can help identify the origins of test-taker confusion (e.g., differences in the understanding of item types, item formats, and other presentation-related aspects of the test) that may lead certain group members to perform differentially on particular items. The interviews can be conducted with test takers from the new populations. Because they can offer rich information about test takers' thought processes as compared to empirical investigations, they can be used as a supplementary method. This approach has been suggested as a cost-efficient way to collect validity evidence by examining potential differences in cognitive processes with small samples of test takers (Ercikan et al., 2010).

Component 3: Generalization

The third component of the framework is generalization. It involves analyzing whether test takers' scores would be comparable across equivalent test forms, which means that a score attained in one testing situation is comparable to scores of other testing instances. This may involve the use of different test forms, administrations, sites, and raters, and in the case of exported assessments, new populations. Procedures to attain such comparability may involve the inclusion of a sufficient number of tasks to provide stable estimates of the performances of the test takers, the configuration of tasks in ways that are appropriate for the desired interpretation (and population), and the use of appropriate scaling and equating procedures. The last two procedures are particularly relevant to international populations.

Duong and von Davier (2014) and von Davier, Holland, and Thayer (2004) discussed considerations for using appropriate scaling procedures based on kernel-equating[2] approaches. These approaches are particularly helpful in instances in which the sample size for at least one of the comparison groups is too small to use the more traditional item-response theory (IRT) models. Alternative approaches include linear equating of test scores on a target population that is composed of multiple groups as well as comparing the results from various calibrations, including those obtained from approaches such as the observed-score equating function (von Davier, 2011) and those obtained from using multigroup calibration in an IRT framework. Such approaches may be helpful in evaluating the appropriateness of the scale, particularly when new populations are included in the testing program. Such analyses may be helpful in evaluating item invariance, which could indicate whether the items are appropriate for the new population. Oliveri and von Davier (2014) suggest as follows:

> . . . when analyzing a test with multiple forms administered to a heterogeneous population, the measurement invariance found in one administration may not be a valid presumption for another administration, which suggests that one sample of test takers may not be exchangeable with another in the case of heterogeneous populations that may vary across administrations, thus confining test results to specific data, and compromising the quality of the reported scores (p. 14).

Explanation, Extrapolation, and Utilization Figure 2 summarizes the definitions, relevant questions, and strategies for explanation, extrapolation, and utilization. These three components are relevant to test-use considerations.

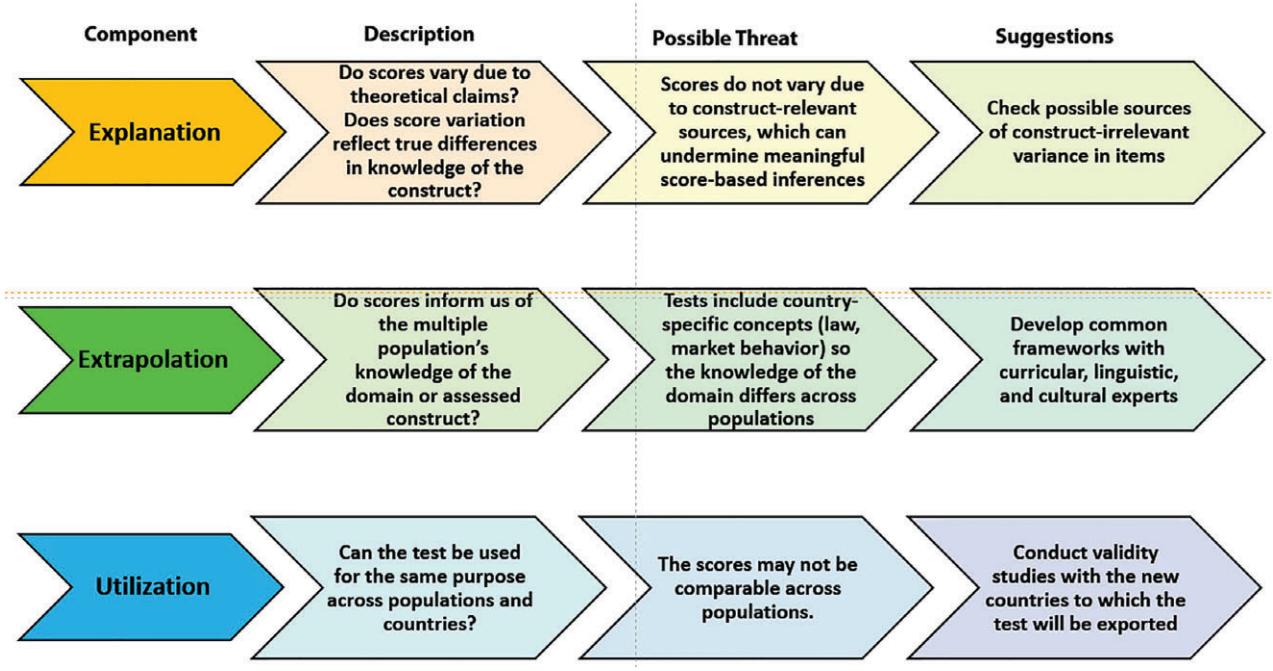

Figure 2 The definition, relevant questions, and strategies for explanation, extrapolation, and utilization components of an assessment that is exported.

Component 4: Explanation

A central notion of the explanation component is whether the variability in scores is primarily related to the construct that is intended to be measured. That is, *Do the scores reflect the cognitive processes, skills, or abilities associated with the assessed construct, and do they vary relative to the degree to which the test takers possess knowledge or skills of the construct*? A high score on a particular measure should indicate that the test taker has high levels of proficiency in the construct the test is intended to measure, as would be expected by theory. In an exported assessment, the evaluation of this assumption would involve minimizing possible sources of construct-irrelevant variance at various stages of test development and test administration so that one can confidently interpret performance differences on the exported assessment in terms of test-taker proficiency rather than on construct-irrelevant differences. This evaluation may involve checking for differences that, for example, may occur due to variations in raters' scoring approaches or their differential use of the rating criteria for the original and new populations.

It should be noted that although elements of this component were described earlier, they were described in relation to the conceptualization of the construct, whereas the description provided in this component is related to analyzing the way the test takers interact with the items after they are developed. For example, an item may be designed to assess a test taker's ability to write a persuasive essay about the benefits of a certain type of medical research. The rubrics for the essay describes the features of persuasive argumentation at the various score points in the scale. However, if a rater does not necessarily agree with the perspective presented by the test taker from the new population because it presents a unique or novel point of view, the rater may attend more to the argument presented, pay less attention to the features of good argumentation as presented in the rubric, and score the essay more harshly. In this case, the rating criteria would not be applied evenly between the original and new populations because of differing points of view and not because of essay quality.

Component 5: Extrapolation

In extrapolation, one aims to extend the inferences made from test scores to proficiency over an entire domain. As Kane (2013) described, the extrapolation component extends the test score interpretation to "real-world" performance and makes a more ambitious inferential leap into a larger and fuller range of performance in the targeted domain to include performances in nontest situations and nontest contexts (Bachman & Palmer, 2010; Kane, 2006). Analyses may involve

examining the degree to which the assessed curriculum is aligned to the test for the multiple populations. A relevant question would thus be: *To what extent is the exported assessment designed to assess curricula and skills that are pertinent to the needs of the populations to which the exported assessment is administered?* The goal of this question is to help minimize construct-irrelevant variance due to differential curricular exposure, which may lead to differential explanations of score variability among test takers of similar ability. That is, test takers may have similar overall abilities on the assessed construct (e.g., quantitative reasoning) but may have a differential understanding of more discrete aspects of the assessed curricula (e.g., statistics and data interpretation). These differences may be due to variations in exposure to aspects of the construct assessed by the test. They also may be due to possible differences in opportunity to learn because of a different emphasis (inclusion or exclusion) of a particular curricular component in the curriculum of one country as compared to another (Oliveri et al., 2018).

Ling (2013) pointed to curricular differences, which may impact performance and interpretations of score differences for non-U.S. versus U.S.-based programs when an assessment is given outside the United States. Ling (2013) also demonstrated that regional differences may appear at the program level. The extent to which the program-level variations of non-U.S.–based programs resemble those of the U.S. programs remains a question worthy of investigation as one considers exporting assessments.

Component 6: Utilization

The utilization component entails implementing the assessment in a way that provides benefits to score users and requires that score interpretations are backed by the highest probability that appropriate and correct decisions will be made from the assessment scores. In the case of exported assessments, it asks whether the assessment has the same or similar predictive power across the original and new countries in which the assessment will be used. Its analysis may involve conducting validity studies across the multiple test-taker populations to have confidence in the level of the predictive validity across populations. As an example, Liu, Klieger, Bochenek, Holtzman, and Xu (2016) conducted a study with a university in Singapore to examine how an admissions test (the GRE General Test) is used in graduate admissions. Findings suggested that the GRE is used differently in Singaporean universities as compared to some institutions in the United States. In Singaporean universities, it is not required of all applicants and is not weighed as heavily as other admissions information. When scores from the GRE General Test are required, it typically is required for Singaporean applicants who did not graduate from one of the top three Singapore universities, suggesting similar uses despite slight variations in utilization across U.S. and non-U.S. contexts.

In a European context, Schwager, Hülsheger, Bridgeman, and Lang (2015) investigated the validity of the GRE General Test in a Netherlands university. The researchers found that the GRE General Test can help predict a graduate student's cumulative graduate GPA and thesis grade for their master's degree, facilitating the comparison of test takers from a variety of backgrounds as part of the admissions decision process. However, because their study was conducted only in the Netherlands, they were unable to generalize the findings to test takers from other European countries. These types of studies can guide future predictive validity studies of other exported assessments.

Conclusion

In this paper, we discussed the challenges associated with exporting assessments using two case-use scenarios as applied examples in higher education: tests used for admissions and SLO assessments. We suggested that there are multiple challenges associated with exporting assessments that include addressing possible sources of construct-irrelevant variance (e.g., differences in opportunity to learn, curricular exposure, and differential understanding of linguistic or geocentric terms sometimes used in assessments). We provided suggestions for ways to reduce this variance to help increase confidence in the use of exported assessments. We reiterate that although we provided a variety of examples of best practices, not all of them need to be implemented. Instead, test developers, researchers, and advisory panel members need to judiciously select which suggestions to implement. Such decisions can be made in light of the context and nature of the testing program and of the various constraints and costs associated with their implementation, as well as the assessment purpose. Constraints may emerge in relation to the resources available to exporting assessments. For example, it might be difficult to recruit adequate numbers of test takers from the targeted (new) population to effectively pilot or

field test items. In this case, cognitive interviews may be one of the most useful sources of information about test takers' interactions with the assessment.

Nonetheless, we suggest considering the range of options described in this paper as a way to increase confidence in the use of exported assessments the world over. Because exported assessments are often used for high-stakes decision-making, an investigation into the existence of such features is particularly important. This paper provides a framework to guide future research to examine the validity of exported assessments. Such studies typically focus primarily on investigating test translation and adaptation as the primary source of test validation; however, as we outlined in this paper, that there are other steps to consider in exportation.

Acknowledgments

The authors would like to thank all members of the working group concerned with this project: Alan Shaw, Sydell Carlton, Bill Sims, Luis Saldivia, and Douglas Baldwin for their help in reviewing the evidence for higher education assessment programs and for the suggestions provided for continuing to maintain and enhance the validity of score-based inferences made from higher education assessments. We also thank Mary Pitoniak, Don Powers, Cathy Wendler, Michael Kane, and Frederic Robin for their feedback and reviews of this report and Hillary Molloy and Kri Burkander for research assistance.

Notes

1 DIF occurs when examinees in one group select a correct item response option at different rates than equally able members of the other group, which may lead to item-level performance differences for one group as compared to the other. DIF may occur due to construct-relevant or construct-irrelevant reasons. Therefore, an analysis of its sources needs to be conducted to determine whether DIF is due to bias (AERA, APA, & NCME, 2014).

2 Equating is a process for relating scores on alternate test forms in order for them to have essentially the same meaning. Equated scores are typically reported on a common score scale (AERA, APA, & NCME, 2014).

References

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

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford, England: Oxford University Press.

Baxter, G. P., & Glaser, R. (1998). Investigating the cognitive complexity of science assessments. *Educational Measurement: Issues and Practice*, *17*, 37–45. https://doi.org/10.1111/j.1745-3992.1998.tb00627.x

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education*, *25*, 27–40. https://doi.org/10.1080/08957347.2012.635502

Chapelle, C. A. (2008). The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–352). New York, NY: Routledge.

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*, 3–13. https://doi.org/10.1111/j.1745-3992.2009.00165.x

Cheung, F. M. (2004). Use of western and indigenously developed personality tests in Asia. *Applied Psychology*, *53*, 173–191. https://doi.org/10.1111/j.1464-0597.2004.00167.x

Cheung, F. M., Leung, K., Fan, R. M., Song, W. Z., Zhang, J. X., & Zhang, J. P. (1996). Development of the Chinese personality assessment inventory. *Journal of Cross-Cultural Psychology*, *27*, 181–199. https://doi.org/10.1177/0022022196272003

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*, 355–368. https://doi.org/10.1111/j.1745-3984.1986.tb00255.x

Duong, M. Q., & von Davier, A. A. (2014). Heterogeneous populations and multistage test design. In R. E. Millsap, L. A. van der Ark, D. M. Bolt, & C. M. Woods (Eds.), *New developments in quantitative psychology. Presentations from the 77th Annual Psychometric Society Meeting* (pp. 151–170). New York, NY: Springer-Verlag.

Dwyer, C. A., Gallagher, A., Levin, J., & Morley, M. E. (2003). *What is quantitative reasoning? Defining the construct for assessment purposes* (ETS Research Report No. RR-03-30). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2003.tb01922.x

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing, 25,* 155 – 185. https://doi.org/10.1177/0265532207086780

Educational Testing Service. (2014a). *ETS® Major Field Test for bachelor's degree in business now available worldwide* [Press release]. Retrieved from http://news.ets.org/news/ets-major-field-test-for-bachelors-degree-in-business-now-available-worldwi

Educational Testing Service. (2014b). *ETS standards for quality and fairness*. Princeton, NJ: Author.

Educational Testing Service. (2014c). *A snapshot of the individuals who took the GRE revised General Test*. Princeton, NJ: Author.

Ercikan, K., Arim, R. G., Law, D. M., Lacroix, S., Gagnon, F., & Domene, J. F. (2010). Application of think aloud protocols in examining sources of differential item functioning. *Educational Measurement: Issues and Practice*, *29*, 24 – 35. https://doi.org/10.1111/j.1745-3992.2010.00173.x

Ercikan, K., & Lyons-Thomas, J. (2013). Adapting tests for use in other languages and cultures. In K. F. Geisinger, B. A. Braken, J. F. Carlson, J. C. Hansen, & N. R. Kuncel (Eds.), *APA handbook of testing and assessment in psychology* (pp. 545 – 569). https://doi.org/10.1037/14049-026

Ercikan, K., & Oliveri, M. E. (2013). Is fairness research doing justice? A modest proposal for an alternative validation approach in differential item functioning (DIF) investigations. In M. Chatterji (Ed.), *Validity, fairness and testing of individuals in high stakes decision-making context*, pp. 69 – 86. Bingley, England: Emerald Publishing.

Hambleton, R. K., Merenda, P., & Spielberger, C. (Eds.). (2005). *Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures*. Mahwah, NJ: Erlbaum.

Huff, K. L. (2000). *Evaluating differential item functioning across selected item formats on a large-scale certification examination* (AICPA Technical Report). Retrieved from https://www.aicpa.org/content/dam/aicpa/becomeacpa/cpaexam/psychometricsandscoring/technicalreports/downloadabledocuments/huff-evaluatingdifferential.pdf

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing, 1*(2), 93 – 114. https://doi.org/10.1207/S15327574IJT0102_1

International Test Commission. (2010). *The ITC guidelines for translating and adapting tests*. Retrieved from https://www.intestcom.org/files/guideline_test_adaptation.pdf

International Test Commission. (2018). *The ITC guidelines in support of the fair and valid assessment of linguistically diverse populations*. Manuscript in preparation.

Kane, M. T. (2006). Validation. *Educational Measurement*, *4*(2), 17 – 64.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1 – 73. https://doi.org/10.1111/jedm.12000

Ling, G. (2013). *Repurposing a business learning outcomes assessment to college students outside of the United States: Validity and reliability evidence* (ETS Research Report RR-13-40). https://doi.org/10.1002/j.2333-8504.2013.tb02347.x

Liu, O. L., Klieger, D. M., Bochenek, J. L., Holtzman, S. L., & Xu, J. (2016). *An investigation of the use and predictive validity of scores from the GRE® revised General Test in a Singaporean university* (GRE Research Report No. GRE-16-01). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12095

Maddox, B. (2015). Inside the assessment machine: The life and times of a test item. In M. Hamilton, B. Maddox, & C. Addey (Eds.), *Literacy as numbers* (pp. 129 – 147). Cambridge, England: Cambridge University Press.

Muñiz, J., Hambleton, R. K., & Xing, D. (2001). Small sample studies to detect flaws in item translations. *International Journal of Testing*, *1*, 115 – 135. https://doi.org/10.1207/S15327574IJT0102_2

Oliveri, M. E., Ercikan, K., & Simon, M. (2015). A framework for developing comparable multi-lingual assessments for minority populations: Why context matters. *International Journal of Testing, 15*, 94 – 113.

Oliveri, M. E., Ercikan, K., & Zumbo, B. D. (2013). Analysis of sources of latent class DIF in international assessments. *International Journal of Testing, 13*, 272 – 293. https://doi.org/10.1080/15305058.2012.738266

Oliveri, M. E., Lawless, R. R., & Mislevy, R. J. (under review). Using evidence-centered design to support the development of culturally and linguistically sensitive collaborative problem-solving assessments. Manuscript in preparation.

Oliveri, M. E., Lawless, R. R., Robin, F., & Bridgeman, B. (2018). An exploratory analysis of differential item functioning and its possible sources in a higher education admissions context. *Applied Measurement in Education, 31,* 1 – 16. https://doi.org/10.1080/08957347.2017.1391258

Oliveri, M. E., Lawless, R. R., & Young, J. W. (2015). A validity framework for the use and development of exported assessments. *Fairness* [Web page]. Princeton, NJ: ETS Office of Professional Standards Compliance. Retrieved from https://www.ets.org/about/fairness/

Oliveri, M. E., & von Davier, A. A. (2016). Psychometrics in support of a valid assessment of linguistic minorities: Implications for the test and sampling designs. *International Journal of Testing, 16*, 205 – 219. https://doi.org/10.1080/15305058.2015.1099534

Oliveri, M. E., & von Davier, M. (2014). Toward increasing fairness in score scale calibrations employed in international large-scale assessments. *International Journal of Testing, 14,* 1 – 21. https://doi.org/10.1080/15305058.2013.825265

Pitoniak, M. J., Young, J. W., Martiniello, M., King, T. C., Buteux, A., & Ginsburgh, M. (2009). *Guidelines for the assessment of English language learners*. Princeton, NJ: Educational Testing Service.

Roth, W.-M., Oliveri, M. E., Sandilands, D., Lyons-Thomas, J., & Ercikan, K. (2013). Investigating sources of differential item functioning using expert think-aloud protocols. *International Journal of Science Education, 35*, 546 – 576. https://doi.org/10.1080/09500693.2012.721572

Schwager, I. T., Hülsheger, U. R., Bridgeman, B., & Lang, J. W. (2015). Graduate student selection: Graduate record examination, socioeconomic status, and undergraduate grade point average as predictors of study success in a western European university. *International Journal of Selection and Assessment, 23*, 71 – 79. https://doi.org/10.1111/ijsa.12096

Shi, L. (2001). Native- and nonnative-speaking EFL teachers' evaluation of Chinese students' English writing. *Language Testing*, *18*, 303 – 325.

Sinharay, S., Dorans, N. J., & Liang, L. (2009). *First language of examinees and its relationship to equating* (ETS Research Report RR-09-05). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2009.tb02162.x

Spielberger, C. D., Moscoso, M. S., & Brunner, T. M. (2005). The cross-cultural assessment of emotional states and personality traits. In R. K. Hambleton, C. D. Spielberger, & P. F. Merenda (Eds.), *Adapting educational and psychological tests for cross-cultural assessment*. Hillsdale, NJ: Erlbaum.

Uysal, H. H. (2008). Tracing the culture behind writing: Rhetorical patterns and bidirectional transfer in L1 and L2 essays of Turkish writers in relation to educational context. *Journal of Second Language Writing, 17*(3), 183 – 207. https://doi.org/10.1016/j.jslw.2007.11.003

von Davier, A. A. (2008). New results on the linear equating methods for the non-equivalent-groups design. *Journal of Educational and Behavioral Statistics*, *33*, 186 – 203. https://doi.org/10.3102/1076998607302633

von Davier, A. A. (2011). A statistical perspective on equating test scores. In A. A. von Davier (Ed.), *Statistical models for test equating, scaling and linking* (pp. 1 – 17). New York, NY: Springer-Verlag. https://doi.org/10.1007/978-0-387-98138-3

von Davier, A. A., Holland, P. W., & Thayer, D. T. (2004). *The kernel method of test equating*. New York, NY: Springer-Verlag. https://doi.org/10.1007/b97446

Wendler, C., & Powers, D. (2009). *What does it mean to repurpose a test?* R&D Connections, 9. Princeton, NJ: Educational Testing Service.

Xi, X. (2010). How do we go about investigating test fairness? *Language Testing*, *27*, 147 – 170. https://doi.org/10.1177/0265532209349465

Young, J. W., Baldwin, D., Chiu, L., Davis, L., Gao, R., Hauck, M., . . . Zieky, M. (2014). *Cultural sensitivity review*. Princeton, NJ: Educational Testing Service.

Young, J. W., So, Y., & Ockey, G. (2013). *Guidelines for best test development practices to ensure validity and fairness for international English language proficiency assessments*. Princeton, NJ: Educational Testing Service.

Zhan, K. (1992). *The strategies of politeness in the Chinese language*. Berkeley: Institute of East Asian Studies, University of California.

Zhang, Y., Dorans, N. J., & Matthews-López, J. L. (2005). *Using DIF dissection method to assess effects of item deletion* (College Board Research Report No. 2005-10). New York, NY: The College Board.

Zwick, R., Thayer, D. T., & Lewis, C. (1999). An empirical Bayes approach to Mantel-Haenszel DIF analysis. *Journal of Educational Measurement, 36*, 1 – 28. https://doi.org/10.1111/j.1745-3984.1999.tb00543.x

Zwick, R., Thayer, D. T., & Lewis, C. (2000). Using loss functions for DIF detection: An empirical Bayes approach. *Journal of Educational and Behavioral Statistics, 25*, 225 – 247. https://doi.org/10.3102/10769986025002225

Suggested citation:

Oliveri, M. E., & Lawless, R. (2018). *The validity of inferences from locally developed assessments administered globally* (Research Report No. RR-18-35). Princeton, NJ: Educational Testing Service. https://doi.org/10.1002/ets2.12221

Action Editor: Donald Powers

Reviewers: Michael Kane, Mary Pitoniak, and Cathy Wendler