Listening. Learning. Leading.®

A Validity Framework for the Use and Development of Exported Assessments

By María Elena Oliveri, René Lawless, and John W. Young

A Validity Framework for the Use and Development of Exported Assessments

María Elena Oliveri, René Lawless, and John W. Young[1]

[1] The authors would like to acknowledge Bob Mislevy, Michael Kane, Kadriye Ercikan, and Maurice Hauck for their valuable input and expertise in reviewing various iterations of the framework as well as Priya Kannan, Don Powers, and James Carlson for their input throughout the technical review process.

Abstract

In this document, we present a framework that outlines the key considerations relevant to the fair development and use of exported assessments. Exported assessments are developed in one country and are used in countries with a population that differs from the one for which the assessment was developed. Examples of these assessments include the *Graduate Record Examinations*® (*GRE*®) and el *Examen de Admisión a Estudios de Posgrado*™ (*EXADEP*™), among others. Exported assessments can be used to make inferences about performance in the exporting country or in the receiving country. To illustrate, the GRE can be administered in India and be used to predict success at a graduate school in the United States, or similarly it can be administered in India to predict success in graduate school at a graduate school in India.

Differences across the multiple populations to which the assessment is administered may include differential understanding of test-taking strategies and behavior, differential understanding of cultural references or idiomatic expressions. Because these differences might be irrelevant to the measured constructs, a framework is needed to ensure score-based inferences are valid and speak about the test takers' abilities rather than their potential lack of familiarity with some aspects of the test that may be construct-irrelevant. To this end, we present our framework, which was inspired by Kane's (2013) validity framework. Kane's framework was used as the lens through which we analyzed validity in exported assessments.

In our framework, we discuss key elements relevant to designing and using exported assessments with multiple populations. We also identify challenges that may need to be faced in order to maintain validity, comparability, and test fairness when using exported assessments and provide recommendations for enhancing the validity of score-based inferences for the multiple populations taking the assessments. These issues are of particular importance, given the growing rates of assessment exportation due to globalization and increased immigration rates among other factors.

Table of Contents

Overview

The number of students studying outside of their homeland is expected to rise from 2.5 million in 2009 to almost 7 million by 2020. A large number of these students are from Asia, entering postsecondary institutions in North America, Western Europe, and Australia (Altbach, Reisberg, & Rumbley, 2009). The use of English as a language of instruction in higher education is also increasing worldwide. A consequence of these trends is the increased use of *exported assessments*. We define exported assessments as ones developed for use with populations in one country and that have new populations added to the test administration through exportation of the assessment to other countries. To illustrate, an assessment that is developed for use in the continental United States may later be used by other English-speaking countries (e.g., Singapore). Likewise, an assessment that may be developed for test takers in Spain later may be marketed and used in other Spanish-speaking countries in South America.

Kane (2013) defined fairness in assessments as the capability of a test to provide scores that have the same meaning across the populations to which it is administered. When assessments are exported, they are administered to multiple populations that may be culturally or linguistically different from the originally intended population. This practice presents complexities in terms of whether the test can yield valid score inferences for the new populations, especially when populations are added after the test was developed.

The emphasis on validity in the context of exported assessments is important because test takers in the new population may possess qualities that may differ from the original populations. These qualities (e.g., tendencies to use different colloquialisms, use of different test-taking strategies based upon culture, differential familiarity with item types or item formats) may impact test performance in construct-irrelevant ways and make the test items less accessible to some examinee groups. Hence, test scores may not be reflective of test takers' abilities but may also be measuring construct-irrelevant factors such as levels of acculturation with the culture where the test was developed.

The presence of construct-irrelevant factors in assessments threaten the validity of score-based inferences if they differentially affect some subgroups, which might lead the resulting scores to have different meanings for the original and the new populations; therefore, exported assessments must ensure that they are assessing the constructs of interest without introducing construct-irrelevant variance to the test.

Examples of Exported Assessments

There are various types of exported assessments. To illustrate, the *GRE*®, *EXADEP*™, and Examen de Ingreso al Posgrado (EXAIP) are examples of high-stakes assessments used for higher education admissions decisions. Another example is the Major Field Test (MFT), which is an assessment used to measure instructional effectiveness.[2] We focus on higher education because the impact of globalization is likely to be greatest at this level. In contrast, K-12 education is typically more influenced by national forces. We will use these assessments as running examples to explicate the various challenges and threats to the validity of exported assessments. We also center our recommendations in the context of these assessments.

One example, the EXADEP, assesses quantitative and analytical reasoning; verbal abilities in Spanish and vocabulary, grammar, and reading comprehension in English as a second language. It was originally developed in 1968 for use in the selection of applicants into higher education institutions in Puerto Rico. It is now administered across multiple Spanish-speaking countries in Central and South America and in Europe (Spain) also for the purpose of admissions. One added use is for awarding scholarships. The total testing volume across these regions accounts for close to 50,000 test takers per year, a nontrivial number (Educational Testing Service [ETS], 2013).

These assessments differ from *multilingual* examinations, which are used for international audiences (e.g., the Programme for International Student Assessment [PISA] or the Progress in International Reading Literacy Study [PIRLS]), which are administered in more than 40 languages. Thus, some of the issues of developing and using multilingual international assessments may be related, but not limited, to challenges in translation and adaptation. A great deal of research has already been conducted on this topic (see Hambleton, Merenda, & Spielberger, 2005 for a review). In contrast, exported assessments are administered in a single

[2] At the time of publication, examples of exported assessments were not found for other contexts, such as in licensure or certification.

language (unless proficiency in a second language is assessed as an additional component as is the case with EXADEP).

Exported assessments also differ from assessments measuring language proficiency such as the *TOEFL*® and *TOEIC*® exams and the TestDaF (Test Deutsch als Fremdsprache or Test of German as a Foreign Language) in which measurement in linguistic proficiency is the primary construct of interest. These types of assessments target test takers' proficiency in a specific language for diverse purposes (e.g., for employment or for admissions into a higher education institution in a particular country). Such assessments were developed for international populations with the intention of assessing linguistic proficiency; therefore, it is within their scope to assess international populations with diverse linguistic and cultural backgrounds. Nonetheless, similar validity issues may arise within such assessments in an attempt to make test items reflect the construct to be assessed and ensure they are devoid of language or geocentric contexts that limit access to understanding the items by the new populations.

Exported assessments are designed to measure constructs other than proficiency in a foreign language. These include (but are not limited to) the measurement of quantitative or verbal skills for a targeted population (e.g., U.S. test takers) and are later marketed to be administered to other populations. This practice raises the question of whether the linguistic or cultural context presented in such assessments is appropriate for the newly targeted populations. The fair and valid use of exported assessments with multiple populations thus requires due diligence on the part of the assessment developer as well as the score users. Thus, in this paper, we describe the considerations that need to be taken into account *prior to* exporting an assessment and using it with multiple populations as a way to derive valid score-based inferences.

Purpose of This Document

Our objectives in this document are threefold. First, we outline the considerations needed to ensure that the development and use of exported assessments can yield valid score-based inferences. Second, we identify challenges that may need to be faced in order to maintain validity, comparability, and test fairness. Third, we provide recommendations for the development of new tests that are valid for multiple populations.

Given the growing rates of assessment exportation mentioned in the introduction of this document, these issues are of particular relevance.

To address these issues, we developed a framework that was inspired by Kane's (2013) validity framework. In the framework, we identify ways in which to evaluate, quantify, and minimize sources of construct-irrelevant variance. We thus aim to increase test validity and fairness in exported assessments and describe ways to ensure that the scores from exported assessments are being used in valid ways.

A Framework for Developing Valid Exported Assessments

Our proposed framework can be conceptualized as a chain of inferences made in the context of exported tests that begin with the scrutiny of the domain to be evaluated (and the construct[s] of interest) and end by examining how well the inferences made from the test scores hold up for the new population(s). We build the framework upon research discussing fairness in the more specific context of administering assessments developed for one population to other (new/multiple) populations. For example, Wendler and Powers (2009) described threats to validity potentially arising in the case where a test is used for a purpose and audience that differ from the ones for which the test was developed. The authors stated that such uses may set limits on the inferences that can be made from assessment scores and suggest two steps to support the short- and long-term use of the test and the interpretation that can be derived from scores: (a) the development of a plausible argument as to why the test should function as expected, and (b) the collection of evidence to support score inferences. We build on this work by specifying the kinds of evidence that need to be collected and the procedures that need to be undertaken in developing fair and valid exported assessments.

We also build upon previous standards, guidelines, and research that describe the importance of developing valid assessments with linguistically and culturally diverse populations. Such documents include guidelines developed by ETS (2009, 2015), Pitoniak et al. (2009), and International Test Commission [ITC] (2005, 2013) as well as the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], and the National Council on Measurement in Education [NCME], 2014). These publications suggest that potential threats to validity may arise in the assessment of diverse test taker groups such as those defined by race, ethnicity, gender, disability, and others due to differential familiarity with item types, content, or vocabulary that may systematically favor one group over another. To maintain and ensure validity in assessments administered to diverse test-taker groups, these publications suggest various approaches including using technically adequate assessment construction procedures to ensure that the assessments are valid for the new populations. These publications set up guidelines and standards related to how to best assess multiple populations, which we use as a guide in our framework. However, these publication s do not provide guidelines that are specific to exported assessments; hence, the critical need for our framework.

Moreover, the framework proposed by von Davier and Oliveri (2013) is relevant. It describes psychometric considerations that can be implemented to take population heterogeneity into account when designing and developing valid assessments for linguistic minorities. Our framework is further guided by Kane (2013) and is in agreement with his definition of validity, which explicates that consistency of score meaning across examinee groups is central to deriving similarly valid conclusions or inferences across the multiple populations to which the test is administered.

Framework Components and Organization

We organize our framework by considering the six components described in Kane's (2013) validity argument to evaluate test fairness. The components are: (a) domain definition, (b) evaluation, (c) generalization, (d) explanation, (e) extrapolation, and (f) utilization; and they are relevant as they elicit considerations of how different aspects of an assessment may impact various test-taker groups. Previously, these components have been exemplified in research in the context of language-proficiency assessments such as TOEFL (Chapelle, 2008; Chapelle, Enright, & Jamieson, 2010; Xi, 2010). This is the first time they are demonstrated in the context of using exported assessments.

We illustrate these components in Figure 1 and demonstrate their interconnectivity. Test takers are placed in the center of the figure to illustrate the importance of keeping them in focus at each step of the interpretative/use argument (IUA) and emphasize that test takers are impacted by every step in building an argument for the use of exported tests. Kane (2013) defined the IUA as capturing the reasoning involved in using test scores for decision-making, which can be evaluated for consistency and plausibility and includes all of the claims based on the test scores, inferences, and assumptions inherent in the proposed use of an assessment and its interpretations.

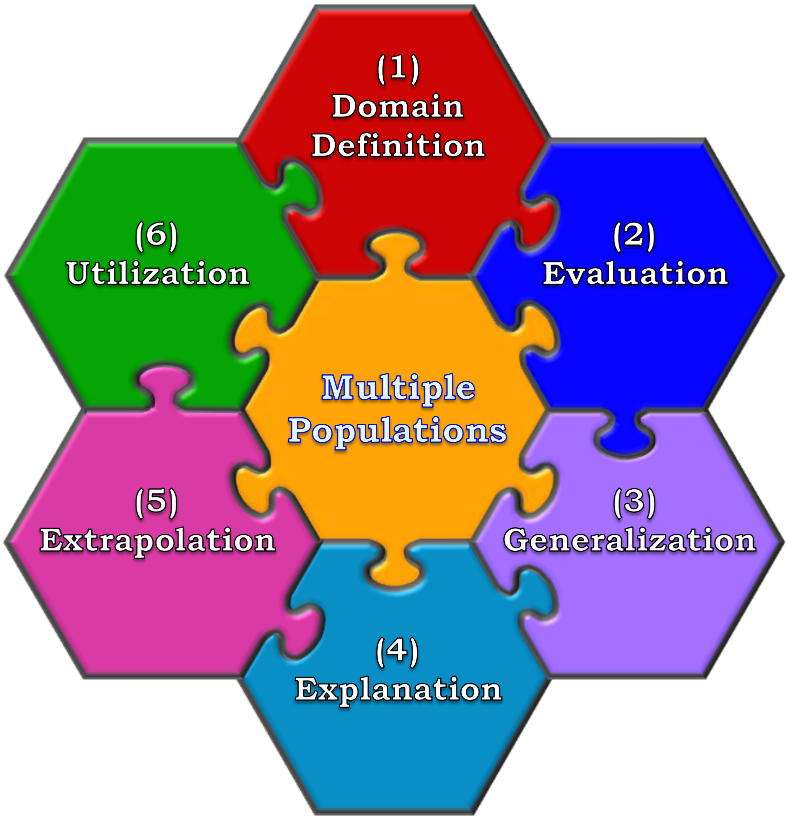


Figure 1. The six components of the framework for the valid use and development of exported assessments.

Next, we describe each of the six components and provide a Toulmin diagram (Toulmin, 1958) to further explain each. We use Toulmin diagrams because they allow us to illustrate the kind of evidentiary thinking one needs to employ in order to make claims for each of the components in our framework. Moreover, the diagrams allow us to describe the steps necessary to ensure a thorough analysis and implementation of the components making up each step to ensure fair exported assessments. Diagrams also provide a practical way to shape and understand the arguments that we make about the fairness of exported assessments. Further, they allow us to examine and evaluate the inferences underlying the IUA systematically, starting from the observed performances and proceeding through the proposed interpretation. Last, diagrams help us identify potential flaws in the argument that might lead to alternative explanations or present conditions for rebuttal against the use of exported assessments for multiple populations. We also provide examples to support the various issues we raise throughout the document.

To understand a Toulmin diagram (see Figure 2 as an example), we begin on the bottom with the *grounds*, which provide the basis for the argument for what we are trying to *claim* about something (found at the top of the diagram). Following the arrows, the *warrant* represents the chain of reasoning that bridges the logic between the grounds and the claim we are trying to make for a particular component of our framework. Behind the warrant are any *assumptions* that need to be made and *backing* statements necessary to support the warrant. On the right side of the diagram (opposite the warrant) are the *rebuttals*, which provide circumstances or counter-arguments under which the claim will not be true, thus reasons for not supporting the inferences about the claims.

In what follows, we explain each of the six components and define the arguments underlying each, with the goal of identifying the underlying attributes to the component, which need to be evaluated prior to exporting an assessment. We also use examples to illustrate relevant issues for each component and enable the reader to consider test validity in the context of ways that minimize construct-irrelevant variance, allowing for enhanced interpretations and uses of the scores for the multiple populations taking the assessment. These issues are of special importance when using exported assessments to ensure comparability and validity of score-based inferences for all test takers throughout the entire testing process.

Component 1: Defining the Domain

Analyzing the domain—within the framework. The first component in the framework for newly exported assessments is defining the domain of interest, which entails identifying the targeted construct(s) to be assessed and the issues that may lessen our ability to measure the targeted construct. This component is illustrated in Figure 2. In defining the domain, the claim is that test takers' performances provide valid evidence of the assessed construct without the introduction of sources of construct-irrelevant variance.

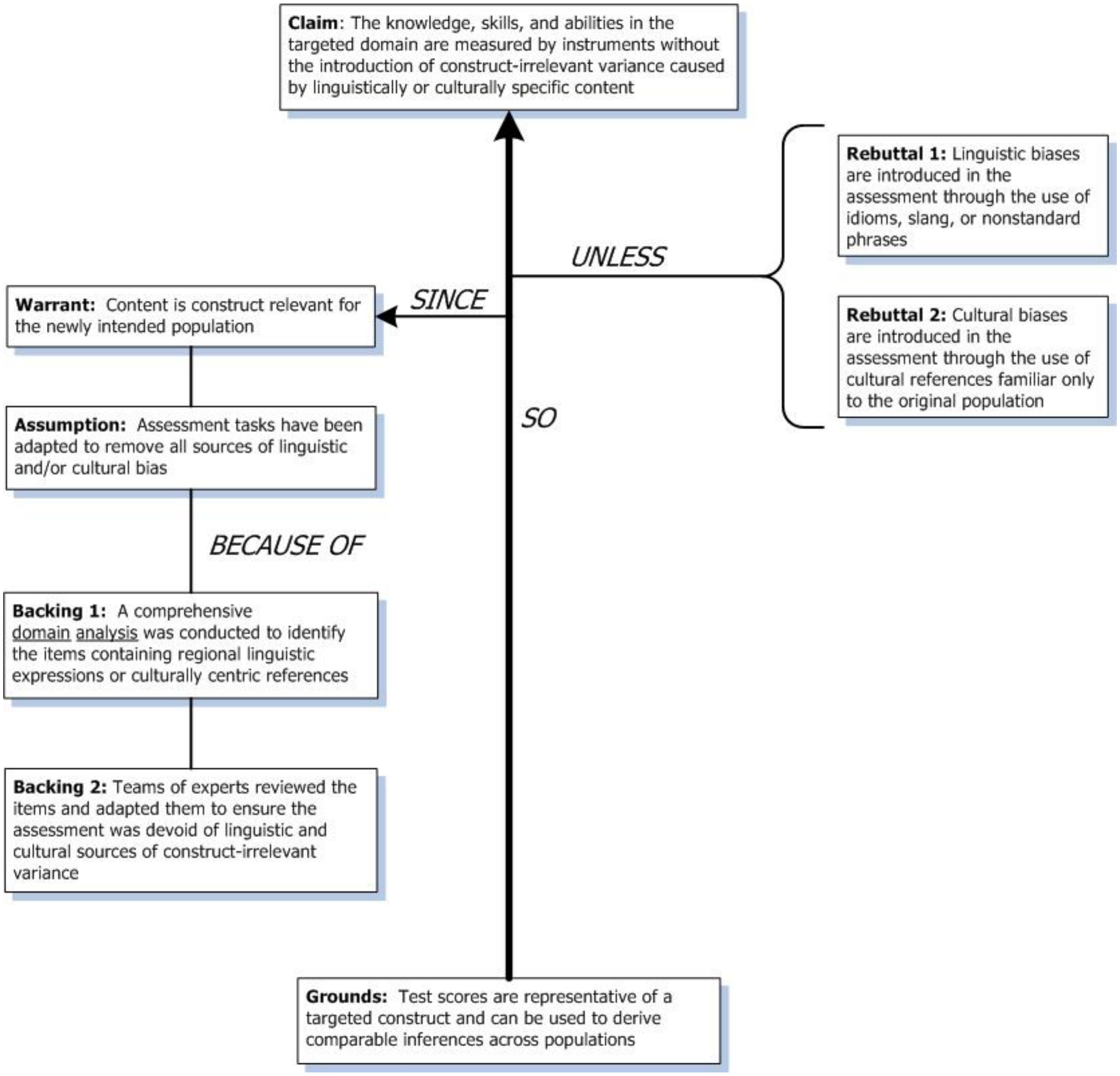

Claim: The knowledge, skills, and abilities in the targeted domain are measured by instruments without the introduction of construct-irrelevant variance caused by linguistically or culturally specific content
Rebuttal 1: Linguistic biases are introduced in the assessment through the use of idioms, slang, or nonstandard phrases
UNLESS
Rebuttal 2: Cultural biases are introduced in the assessment through the use of cultural references familiar only to the original population
Warrant: Content is construct relevant for the newly intended population
SINCE
SO
Assumption: Assessment tasks have been adapted to remove all sources of linguistic and/or cultural bias
BECAUSE OF
Backing 1: A comprehensive domain analysis was conducted to identify the items containing regional linguistic expressions or culturally centric references
Backing 2: Teams of experts reviewed the items and adapted them to ensure the assessment was devoid of linguistic and cultural sources of construct-irrelevant variance
Grounds: Test scores are representative of a targeted construct and can be used to derive comparable inferences across populations

Figure 2. Toulmin diagram for domain definition.

The warrant in defining the domain posits that assessments administered to multiple populations are construct-relevant for the examined domain. The presence of construct-irrelevant factors undermines our ability to make valid score-based inferences and is important because tests containing high cultural loads, meaning that items on the test require “specific knowledge of, or experience with, mainstream … culture” (Rhodes, Ochoa, & Ortiz, 2005, p. 186), also may be assessing test takers’ level of acculturation or “learning of the culture(s) in which the person is expected to demonstrate competence” (Helms, 1992, as cited in Helms, 1997) in addition to the assessed construct. Given that sources of construct-irrelevant variance threaten test fairness, the goal in the review of a test should be to pare down the task demands to construct-relevant ones. In the case of exported assessments, it is important to strive to attain this goal to maintain the validity of the inferences derived from assessment scores (described in another component). To this end, we suggest thinking deliberately at early stages of test review about the construct(s) of interest in terms of the different aspects of validity as well as about the potential sources of construct-irrelevant variance. One can then connect the inferences derived from assessment scores to the interpretation of the resulting test scores and also identify potential issues that might limit the confidence with which one can use test scores to derive valid-score based inferences for the multiple populations taking the assessment.

We suggest using an evidence-centered design (ECD: Mislevy & Haertel, 2006; Mislevy, Steinberg, & Almond, 1999) approach for creating a clear definition of the key knowledge, skills, and abilities (KSAs) to be assessed. The use of ECD allows one to address issues regarding a test’s existing design while simultaneously mapping out evidence of how test takers’ performances support the inferences that are made regarding the measured construct. It also allows one to consider the potential sources of construct-irrelevant variance that are inadvertently introduced in a test when it is administered to multiple populations. For example, an assessment containing novel item types (e.g., complex multiple choice, fill in the blanks) may lead to differential performances for the new population due to construct-irrelevant reasons.

An example of linguistic differences is the differential degree of familiarity with particular terms (such as idioms) contained in the language of the test. We illustrate this issue using English as an example. In this case, although members of populations from two different countries are speakers of English, there may be linguistic (dialectic) differences that need to be considered during test review. The use of idioms or phrasal verbs in test material could disadvantage an entire population, even though the residents of both countries speak the same language. As an example, in British English, the use of the idiom *on the blower*[3] would be completely foreign to an examinee from the United States. Similarly, the American English idiom *wet blanket*[4] may be foreign to an examinee from Great Britain. So although members of the populations from both countries are native speakers of English, there are linguistic (dialectic) differences that need to be considered during test development or item reviews as they disadvantage the new population, even though test takers from both countries speak the same language.

These issues are well-documented in various standards and guidelines. For example, as indicated in the ITC guidelines (2013), test developers/publishers should ensure that tests take into full account the linguistic and cultural differences among members of the multiple populations taking the assessment. Similarly, as indicated by the *Guidelines for the Assessment of English Language Learners* (Pitoniak et al., 2009), widely accessible vocabulary should be used and colloquial and idiomatic expressions or polysemous vocabulary (words with multiple meanings) and unduly challenging words that are not part of the construct should be avoided as is done in the context of assessing English language learners, which may be similar to when multiple populations are assessed. In relation to vocabulary usage, it also suggests that test developers evaluate a list of specialized vocabulary used on the test to determine whether any of the entries are likely to be a source of construct-irrelevant variance for a population. Similarly, references to particular words or concepts may be used in an assessment (e.g., sports that are predominantly played in specific regions of the world such as ice hockey, played primarily in northern countries) may not be common in all parts of the world.

Another related issue may be relying on the use of particular geocultural contexts or content in item development (e.g., U.S.- or British-centric locations that may be unfamiliar to test takers outside those regions or countries). For example, in the context of driving, some drivers may be familiar with references to rotaries while others are familiar with circles. Another example would be the use of references to specific monetary

[3] The idiom *on the blower* means to be on the telephone.
[4] The idiom *wet blanket* refers to someone who lacks enthusiasm about something and tries to prevent others from having a good time.

currencies, as they differ across countries. The reliance on the use of geocentric content may make exporting a test more difficult if the items may be too country- (population-) focused, for example, making references to the 50 states in the United States or the provinces of Canada. This is why the *ETS International Principles for Fairness Review of Assessments* (ETS, 2009) suggested that item context must be adapted for the culture of the countries in which the items will be used when multiple populations are involved in taking the assessment. As a remedy, a comprehensive domain analysis (Backing 1) should be conducted using teams of experts (Backing 2) to evaluate the claim that items covering the assessed domain are representative of the examined construct and are similarly accessible to the diverse populations taking the assessment (the issue of accessibility is further discussed in AERA, APA, & NCME, 2014, pp. 52–53).

Analyzing the domain—the role of experts. The practice of conducting a comprehensive domain analysis involves making certain that an assessment contains only domain-specific information. Otherwise the claim, warrant, and grounds may be invalidated or weakened and the test scores may not be representative of the target construct for the multiple populations.

An important step in this evaluation is conducting reviews with experts to ensure that the test items, test instructions, and stimuli do not contain information that is construct-irrelevant. The expert reviews could consist of a group review with a discussion or an individual expert review of the items in the assessment. The intention of expert reviews should be to examine potential differences in the way constructs are assessed for different populations. This type of review ensures that the test measures the construct as intended and does not contain high linguistic or cultural loadings that may disadvantage the new population. With reference to the multiple populations, it involves asking questions such as the following: Is there language included in the item that is unnecessary, too complex, or extraneous for a test taker to understand and respond to? Is the vocabulary in the items part of the assessed construct? Are the items contextualized in situations unfamiliar to some populations? The results of these reviews should then be shared with test developers so that they can remove potential sources of construct-irrelevant variance from the items.

Expert reviews need to be carefully planned and implemented by selecting experts familiar with the original test population and targeted population. They should also be familiar with the nuances of the language of the test, able to recognize the features of the language that may be problematic, and be sensitive to text that may be specific to a particular culture to ensure that the tests are fair for the new population(s). Materials should be prepared that give examples of language that may be problematic if they appear in the items. If individuals are conducting the reviews without discussion, it is recommended that items be reviewed by at least two experts so that test developers pay close attention to the adaptation of those items so as to avoid clearly problematic language (Bracken & Barona, 1991; Brislin, 1980, 1986; Geisinger, 1994; Greenfield, 1997; Hambleton, 1994; van de Vijver & Hambleton, 1996; van de Vijver & Leung, 1997a, 1997b; van de Vijver & Tanzer, 1997). An implicit assumption of test scores is that test takers who respond correctly to the test items possess the required ability or understanding of the assessed construct and their incorrect responses do not reflect linguistic or cultural biases contained within the test. This could occur when the language or dialect used by the test taker and the test differ (Laing & Kami, 2003) or due to differences in the context of the test items and the background of the test takers. As suggested by Rhodes et al. (2005):

> In order to understand the functioning of an individual on a measured task, we must first understand the influences which caused the individual to perform in the manner observed. When we fail to account for such culturally based behaviour, we run the greatest risk of identifying simple differences as serious deficits (p. 136).

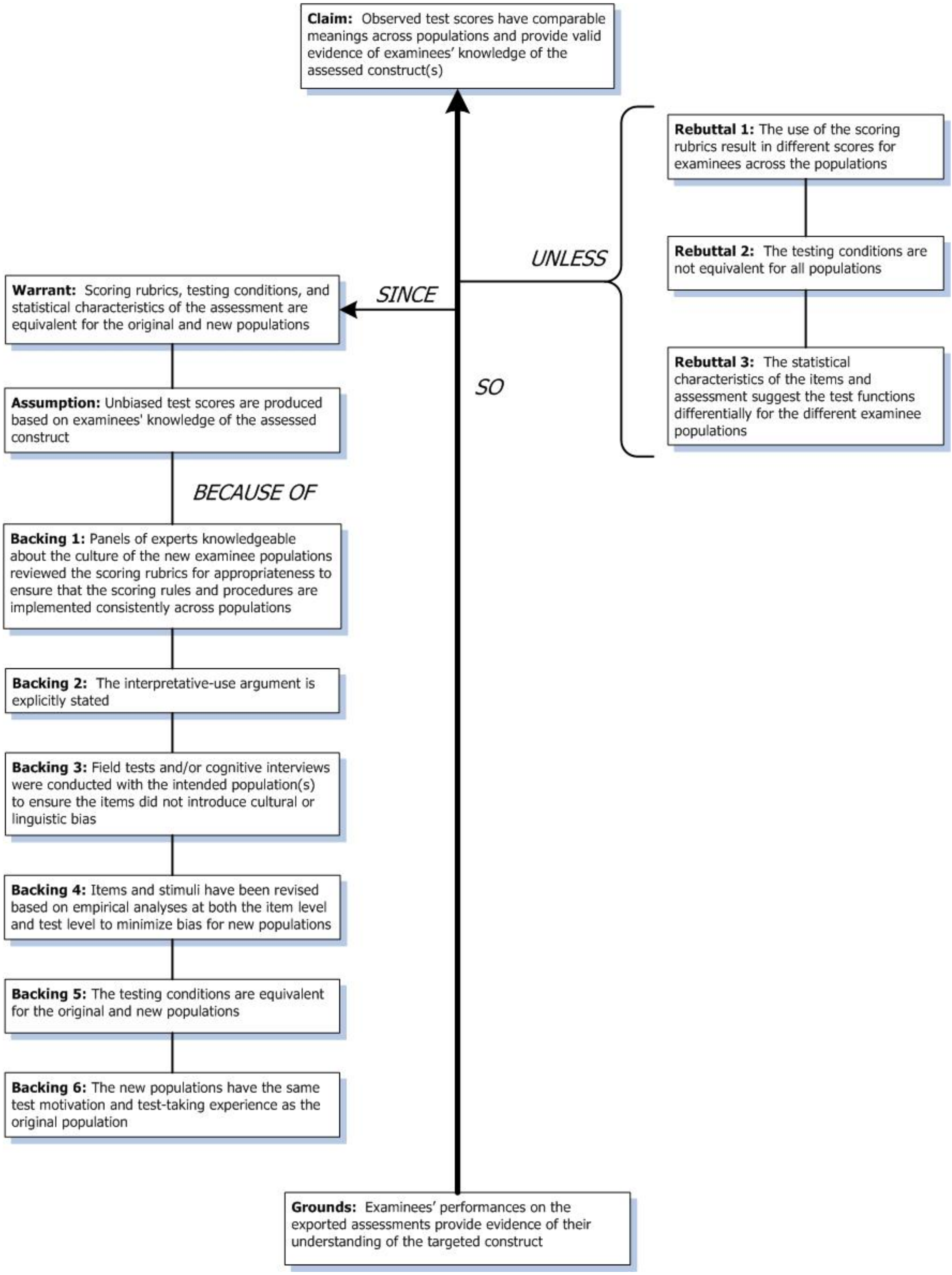
As an example, a mathematics item that is couched in the sport of curling may be conceptually strange and likely unknown to test takers in a Caribbean country but be familiar to test takers in northern countries. Incorrect responses, therefore, from test takers in warmer climates may be due to construct-irrelevant reasons. In fact, the mere context of this sport may cause confusion for these students by focusing their attention on the context, which may be unrelated to the mathematics. This does not necessarily mean that such students do not have the mathematics skills necessary to answer the item; it may mean that they have been confused with what the item is asking them to do. Hence, if a different context had been used, these test takers may have been able to respond correctly and demonstrate their mathematics skills more accurately. This example suggests that interpretation about which test takers have the assessed skill cannot be made without also considering the potential construct-irrelevant variance that may be introduced in the item due to lack of familiarity with the suggested context.

Expert reviewers can also identify differences in the targeted population's familiarity with (a) specific item types, (b) test-taking strategies, and (c) the structure of the assessment in question (e.g., the sequencing of items by difficulty). One way to help mediate individuals' unfamiliarity is to provide them with comprehensive test information and practice materials.

Component 2: Evaluation

The next step in the framework is to evaluate the degree to which the scores derived from a test are plausible and appropriate for its proposed use. Questions at this stage include the following: Are the scoring rubrics designed solely to capture the construct of interest for the multiple populations to which the test will be administered, allowing for valid score-based inferences (as opposed to focusing on the language or mechanics of responses where the construct is not related to language proficiency)? Do the scoring rubrics provide raters with ways to assess the targeted construct(s) for the multiple populations taking the assessment? Are scoring notes created for constructed-response items to inform raters of not penalizing responses for construct-irrelevant reasons (e.g., for penalizing an essay for incorrect grammar when the construct is unrelated to sentence construction)? Do the use of the scores align with the intended and prespecified use of the test?

As summarized in Figure 3, the ability to make claims about a score and its meaning for multiple populations requires that the scores have comparable meanings for the different populations. If there are alternative interpretations for scores across the various populations, test validity may be jeopardized. Hence, the assumptions of score comparability and equality of score meaning need to be evaluated to identify the potential aspects of the rubrics that may unfairly disadvantage a test-taker population. For example, a scoring rubric that includes "succinct writing" as an aspect of good writing may disadvantage members of a population who believe that it is impolite to be direct and consider writing in a less direct way as more appropriate.



Claim: Observed test scores have comparable meanings across populations and provide valid evidence of examinees' knowledge of the assessed construct(s)
Rebuttal 1: The use of the scoring rubrics result in different scores for examinees across the populations
Rebuttal 2: The testing conditions are not equivalent for all populations
Rebuttal 3: The statistical characteristics of the items and assessment suggest the test functions differentially for the different examinee populations
UNLESS
SINCE
SO
Warrant: Scoring rubrics, testing conditions, and statistical characteristics of the assessment are equivalent for the original and new populations
Assumption: Unbiased test scores are produced based on examinees' knowledge of the assessed construct
BECAUSE OF
Backing 1: Panels of experts knowledgeable about the culture of the new examinee populations reviewed the scoring rubrics for appropriateness to ensure that the scoring rules and procedures are implemented consistently across populations
Backing 2: The interpretative-use argument is explicitly stated
Backing 3: Field tests and/or cognitive interviews were conducted with the intended population(s) to ensure the items did not introduce cultural or linguistic bias
Backing 4: Items and stimuli have been revised based on empirical analyses at both the item level and test level to minimize bias for new populations
Backing 5: The testing conditions are equivalent for the original and new populations
Backing 6: The new populations have the same test motivation and test-taking experience as the original population
Grounds: Examinees' performances on the exported assessments provide evidence of their understanding of the targeted construct

Figure 3. Toulmin diagram for evaluation.

Steps in evaluating validity. We discuss four steps to evaluating validity in tests administered to multiple populations. These are conducting field tests, analyzing field test data and the presence of differential item functioning (DIF), conducting cognitive interviews, and using experts to evaluate the fairness of the rubric. DIF occurs when people who have the same ability on the latent trait measured do not have an equal probability of responding correctly to the test item (AERA, APA, & NCME, 2014). The presence of DIF may suggest potential bias with a test item due to construct-irrelevant sources of variance. Early identification of such items can ensure that they are reviewed and, if necessary, appropriately revised prior to the operational administration of the assessment.

Field testing. To evaluate whether a construct of interest is equivalent across groups and to ensure the common understanding of a construct, field tests should be conducted with the test-taker populations of interest. There are various recommendations for how to conduct field tests. For example, the *Guidelines for the Assessment of English Language Learners* (Pitoniak et. al., 2009) suggested conducting small scale pilot tests with a sample of test takers similar to those who will take an assessment operationally. This was suggested for several reasons including to inform decisions about the appropriateness of test items for the particular sample of test takers, inform content and fairness reviews of the items, and evaluate timing requirements for the different item types. In addition, these studies can also be used to evaluate the clarity of the test instructions. After field testing, interviews can be held with the participants to capture any information that may not be obvious during data analysis.

Analysis of field test results and differential item functioning. The analysis of field test data is an important step in test construction, particularly in relation to administering tests with multiple test-taker populations. Field testing helps identify items that may contain construct-irrelevant factors due to cultural or linguistic differences. For example, it may be discovered during field tests that item wording interferes with the understanding of a test item. Results from field test analyses can inform decisions on whether particular items should be retained or discarded from the assessment (e.g., those items that function differentially across test-taker groups) or modified (i.e., to eliminate problematic wording). Field test results are also useful for monitoring how equating/linking items perform (if these were included in the assessment). Using the field test results, DIF analyses can be conducted if the items are administered to a large enough sample of examines (around 300 from each population) and can detect whether the items function similarly across the intended populations.

Cognitive interviews. Cognitive interviews (also referred to as think-aloud studies) can be particularly useful in identifying any hidden assumptions or alternative plausible interpretations of the test scores. This serves as a useful, additional step toward understanding why some of the items may behave differentially for various populations. Cognitive interviews can help uncover differences in thought patterns between the original and new populations and identify the sources of confusion (for test takers) that may lead the groups to perform differentially on the items. These interviews can also help identify differences in the understanding of item types and formats and other presentation-related aspects of the test that may differentially impact the populations. Further, cognitive interviews can help uncover whether all test takers employ the same test-taking strategies in responding to items . These interviews can be conducted on a representative subset of test takers from the newly targeted populations and can be a cost-efficient way to collect validity evidence by empirically examining potential differences in cognitive processes with small sample sizes (Ercikan et al., 2010).

Evaluating the fairness of the rubrics. Another important step in evaluating validity is to examine the fairness of the rubrics. As shown in Figure 3, the assumption associated with evaluating fairness is that the scoring rubrics, testing conditions, and statistical characteristics support the claim reflecting valid evidence of test takers' knowledge of the assessed construct. The ability to use these scores across populations relies on this assumption. As indicated in the backing statements for scoring inferences, this typically involves obtaining the judgments of expert panels who develop and review the scoring criteria and provide evidence that the scoring procedures are implemented consistently and correctly (Clauser, 2000). This step may resemble a kind of calibration of the expert reviewers. Without this step, it may not be possible to evaluate the validity of the test scores. This is important because score comparability across populations enables one to draw valid inferences about the targeted construct(s), which may include evidence for a wide-range of tasks and contexts. To this end, panels of experts can be used to identify those aspects of the rubrics that are not amenable to assessing multiple populations. In addition, psychometricians can carry out reliability studies to check for scoring consistency for the exported assessments (Kane, 2013).

When considering the use of an assessment with a new population, it is necessary to evaluate the scoring rubrics to ensure that they specify the key elements of a correct response to the item, to allow one to move from the test performance (data) to valid scores (Kane, 2013). The appropriateness of the scoring rubric should be evaluated to ensure that it is not disadvantaging any test-taker population. We suggest addressing this issue in the early stages of test review as new populations are added, to remove potential sources of construct-irrelevant variance. When the IUA is developed, it should be evaluated by a neutral or critical party during the examination of test content by the expert review panels.

These issues are also indicated in the ITC (2013) guidelines, which suggested that test developers/publishers should provide evidence that the language used in the test directions and rubrics are appropriate for all test-taker populations for which the test or instrument is intended. In scorer training for constructed-response items, benchmark training papers that reflect characteristics of the various populations can be useful to familiarize scorers with how the rubric applies to a range of responses. All aspects of scorer training—prior to and during scoring—should include responses from the full range of populations (if they can be identified) as part of the training materials. Recalibrating scorers at the beginning of each scoring session should ensure scorers' abilities to continue scoring accurately. For example, an effective practice may be to examine score discrepancies in the rating of essays from test takers between multiple populations and to conduct checks through back-reading by a chief reader/table leader to settle the discrepancies that do occur between scorers and address score drift with the raters as it occurs through recalibration using benchmark essays (Oliveri, Gundersen-Bryden, & Ercikan, 2012). Various approaches might be used to rating. One is to use the same raters and the same rubrics for the different populations. An advantage of this approach is that it might help minimize score variance due to construct-irrelevant reasons because the use of a different pool of raters might insert an additional challenge associated with calibrating the new and original set of raters. We also suggest supplementing the rubrics with scoring notes to explain cultural or linguistic nuances that are not relevant to the construct being measured such as differences that may occur in writing such as more expressive or succinct essays.

Component 3: Generalization

Assume that the first two domains in the framework have been appropriately addressed. That is, the exported assessment has been sufficiently reviewed for the new population so that the domain and construct have been well-defined and the use of superfluous or confusing language and culturally bound references has been removed. Next, assume that the appropriateness of the rubrics and test use have been evaluated for the new population. Our next line of investigation would concern the reliability of student performances across parallel test forms. We ask whether, after suitable modifications to the assessment have been made for exporting the test, does it stand to reason that the items on the test are still a random sample of items that are representative of the targeted domain (Kane, 2013)?

The generalizability of the scores on an assessment is evaluated after test development. As shown in Figure 4, an evaluation of generalization involves examining whether the configuration of the tasks is appropriate for the intended score interpretations and whether the test has a sufficient number of tasks to demonstrate the test takers' knowledge, skills, and abilities (KSAs) in the construct(s) of interest. Hence, the score obtained on one testing instance should be equivalent to that of other testing instances, which may involve different testing forms, tasks, test sites, test administrations, and where raters are involved, other raters.

A consequence of improper or inadequate training of the raters may limit the interchangeability of test forms across populations. Some test-taker groups might be disadvantaged by a group of raters who had less familiarity with a population's writing style, and so the test takers receive different (lower) scores than if they had been scored by a different set of raters having a greater degree of familiarity with the new population. This issue renders rater training and evaluating reliability and test consistency as critical issues to consider in support of validity. If inconsistencies are identified, the scoring rubrics should be modified and raters should be trained to attend to such issues. We also emphasize that, during pretesting or other forms of item tryout, the sampling of conditions of observation must be extended to the multiple populations. In other words, samples of test takers should be drawn from the various groups to which the test will be administered and efforts must be made to include them in the universe of generalization. We suggest that the universe of items from which to sample should be bias-free for the multiple populations taking the assessment.

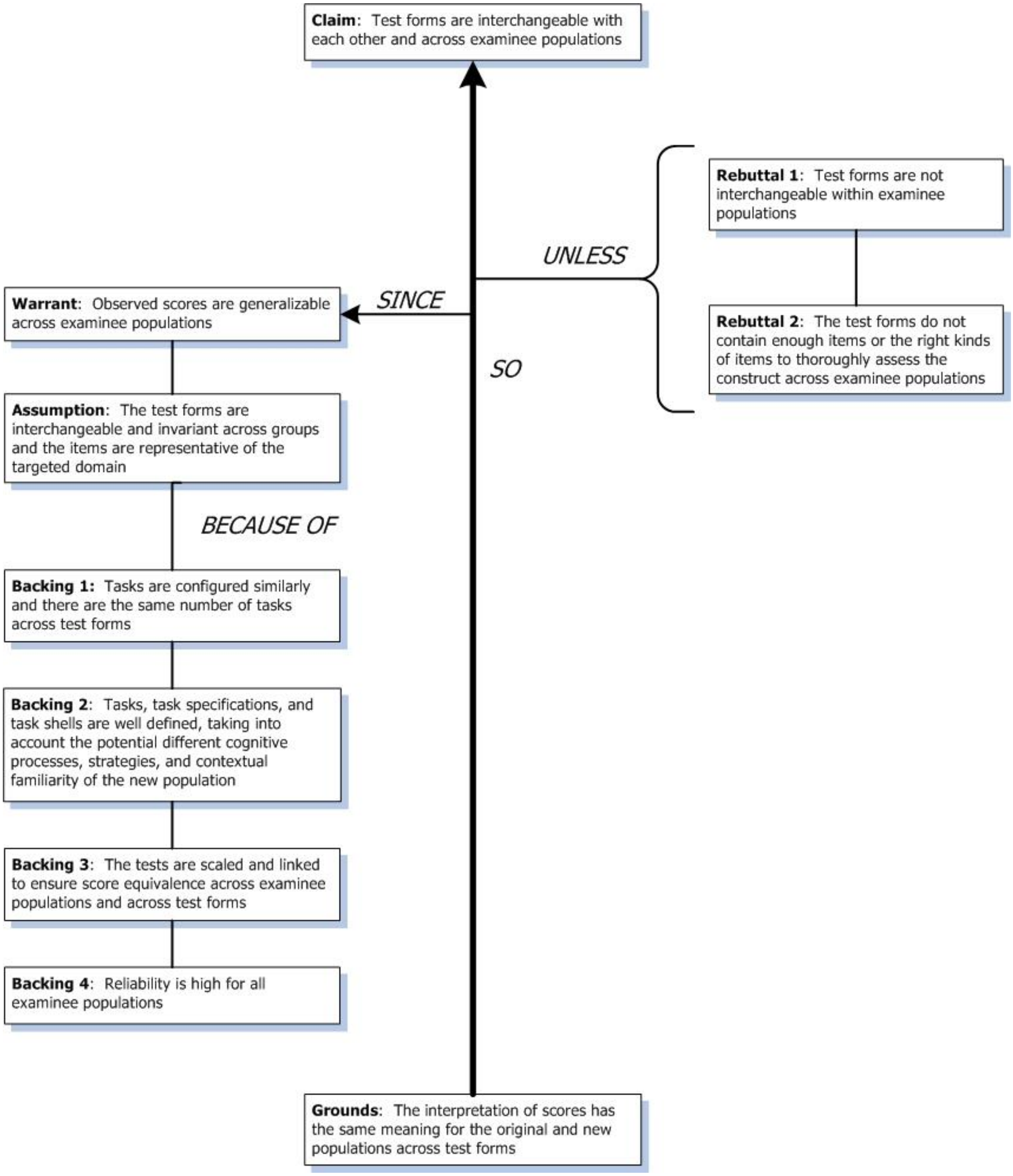

Claim: Test forms are interchangeable with each other and across examinee populations
UNLESS
Rebuttal 1: Test forms are not interchangeable within examinee populations
Rebuttal 2: The test forms do not contain enough items or the right kinds of items to thoroughly assess the construct across examinee populations
SINCE
Warrant: Observed scores are generalizable across examinee populations
SO
Assumption: The test forms are interchangeable and invariant across groups and the items are representative of the targeted domain
BECAUSE OF
Backing 1: Tasks are configured similarly and there are the same number of tasks across test forms
Backing 2: Tasks, task specifications, and task shells are well defined, taking into account the potential different cognitive processes, strategies, and contextual familiarity of the new population
Backing 3: The tests are scaled and linked to ensure score equivalence across examinee populations and across test forms
Backing 4: Reliability is high for all examinee populations
Grounds: The interpretation of scores has the same meaning for the original and new populations across test forms

Figure 4. Toulmin diagram for generalization.

Component 4: Explanation

At the explanation stage, the focus is on ensuring that the cognitive processes and KSAs associated with the assessed construct vary only relative to the degree to which the test takers possess knowledge or skills of the assessed construct. Hence, the inference that can be derived from a high test score is that a high-performing test taker has high levels of KSAs on the assessed construct as would be expected by theory. If a measure is completely free of linguistic or cultural sources of construct-irrelevant variance, the scoring rubrics and scoring mechanisms are found to function sufficiently well for the original and new populations, and the test forms are found to be interchangeable, the assumption would be that the new population would be able to demonstrate the same KSAs on the assessed construct as the original population.–The Toulmin diagram of explanation is found in Figure 5.

One aspect of explanation is to ensure that performances on tests are correlated with performances on other measures. In the context of using exported assessments, it may be important to see how well scores on the new test relate to more localized measures of the same construct. However, it is important also to note that there may be discrepancies between the quality of localized tests and the exported tests; hence, potentially, differences in correlations may be a function of differences in quality. On the other hand, if the localized and exported measures are of similar quality, one could make comparisons between them (e.g., EXADEP versus a local measure used for admissions purposes in a Spanish-speaking country). This step may include examining the degree to which EXADEP scores are related to locally administered assessments (e.g., scores from assessments administered by local universities in that country, if the education system is decentralized, or in comparison to a national examination in the case of a centralized education system). In the context of administering tests to new populations, several issues could jeopardize the inference that low test scores are indicative of test takers having low levels of KSAs on the assessed construct. One is related to the ways in which expertise is developed and how it is demonstrated. Mislevy (2010) explained that expertise is developed and can be demonstrated not only through the recognition of patterns observed in the physical world but also those observed in the social world. Both types of patterns have an influence on how people learn:

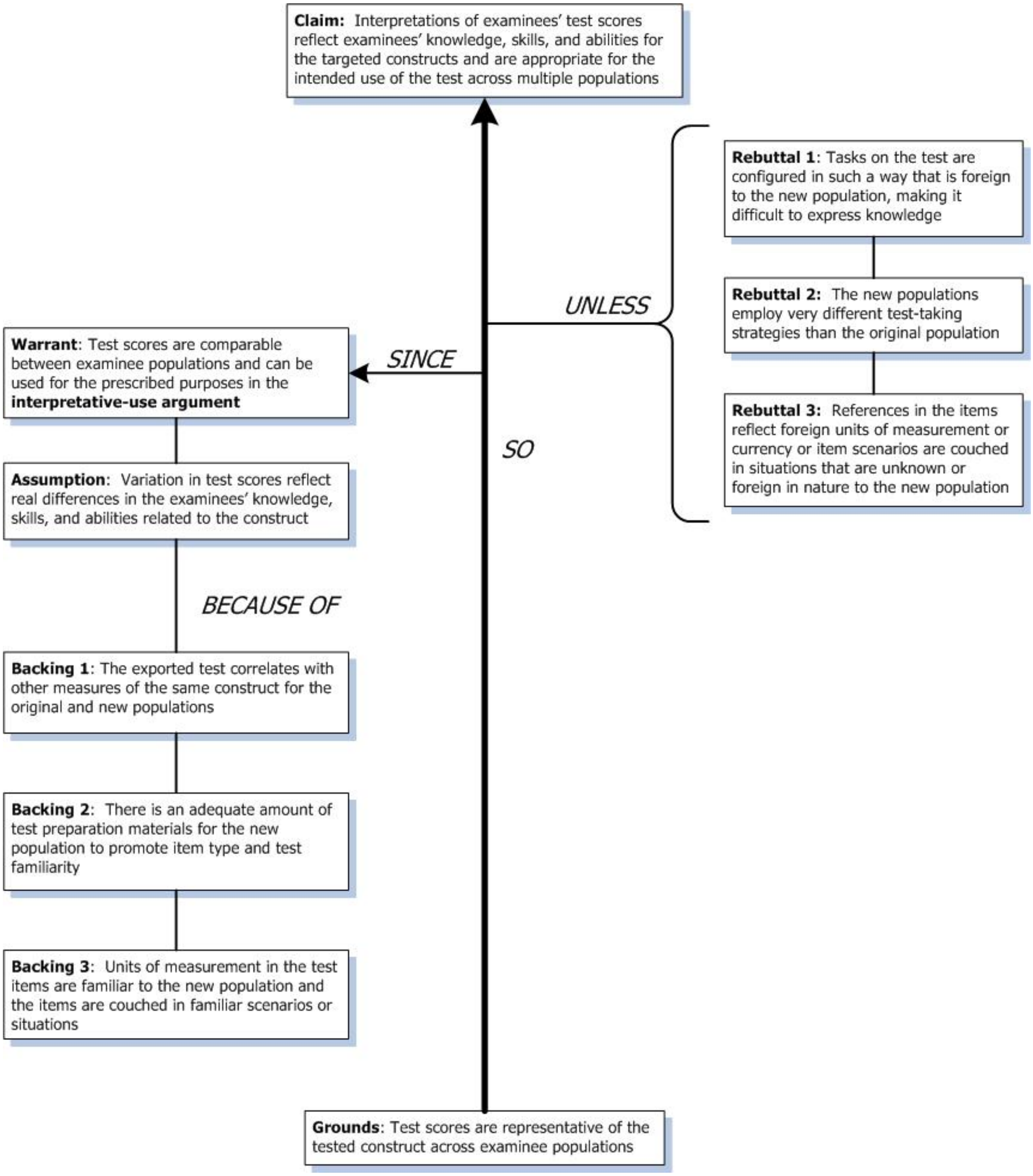

Claim: Interpretations of examinees' test scores reflect examinees' knowledge, skills, and abilities for the targeted constructs and are appropriate for the intended use of the test across multiple populations
UNLESS
Rebuttal 1: Tasks on the test are configured in such a way that is foreign to the new population, making it difficult to express knowledge
Rebuttal 2: The new populations employ very different test-taking strategies than the original population
Rebuttal 3: References in the items reflect foreign units of measurement or currency or item scenarios are couched in situations that are unknown or foreign in nature to the new population
SINCE
Warrant: Test scores are comparable between examinee populations and can be used for the prescribed purposes in the **interpretative-use argument**
SO
Assumption: Variation in test scores reflect real differences in the examinees' knowledge, skills, and abilities related to the construct
BECAUSE OF
Backing 1: The exported test correlates with other measures of the same construct for the original and new populations
Backing 2: There is an adequate amount of test preparation materials for the new population to promote item type and test familiarity
Backing 3: Units of measurement in the test items are familiar to the new population and the items are couched in familiar scenarios or situations
Grounds: Test scores are representative of the tested construct across examinee populations

Figure 5. Toulmin diagram for explanation.

What we look for in situations, how we think about them, how we talk about situations with other people, and how we use the tools and representations that have been developed in a particular domain. Acquiring expertise means capitalizing on the kinds of learning and cognition that one is naturally good at, in order to be able to tackle situations for which one is not naturally good. Although this principle has been applied primarily to cognition and learning, a similar principle could be applied to assessments. This principle suggests that the social environment and the way in which tasks are configured or presented in an assessment context may have a bearing on how knowledge and skills are represented. So if a task is configured in a way that is foreign to the new populations to which the test is administered, such populations may have greater difficulty expressing their knowledge. Such difficulty in knowledge expression may be incorrectly thought of as a lack of knowledge rather than an unfamiliarity with the way in which the information is collected. This may make it more difficult to ensure that the processes and strategies required to successfully complete tasks vary according to theoretical expectations. To remedy such a situation, we suggest producing test preparation materials to help members of the new population familiarize themselves with the requirements of the items. Producing test preparation materials requires a universal approach to thinking about the assessment of the construct. For example, ensuring the use of scenarios or contexts that are broadly applicable may be useful when administering assessments to multiple populations. Similarly, minor adaptations to items such as the use of different units of measurement or contexts that are more familiar to test takers may enhance their ability to demonstrate the KSAs they possess. For example , the use of miles as the unit of measurement in a mathematics item asking about distances may be cognitively more challenging for test takers who are used to the metric system and may lead them to focus on using conversion skills instead of distance calculations. A minor change to the same item contextualizing it in the metric system would allow for such test takers to use their cognitive processes for the demonstration of their KSAs in the construct of interest.

Component 5: Extrapolation

The basis for extrapolation lies in the idea that the interpretations of test scores can be extended to a test taker's proficiency more broadly over an entire domain. Unlike generalization, where the focus is on comparable results between parallel forms and across examinee populations, extrapolation entails generalizing from the test to the domain and its constructs. In other words, a test taker's proficiency is not simply indicative of performance on a specific test but, in fact, is an indicator of how well the test taker will perform on any task that tests the same construct(s) in the same domain. It also presumes that the cognitive processes that are utilized by the test taker will be applicable to other tasks in the same domain. Therefore, the interpretation of a test score would allow for extrapolating a test taker's KSAs over the breadth of the assessed domain. In the case of exported assessments, the use of tests that have not been properly reviewed for linguistic or cultural loadings and pretested on a new populations may diminish the ability to derive test scores that elicit the appropriate evidence from test takers on the targeted construct. For example, an assessment that is developed for use in admissions decision-making for U.S. higher education institutions may not be similarly useful or appropriate for use in admissions to universities in Asia, which may require different skills or cover different curricula.

Figure 6 specifies the considerations for extrapolation. One aspect of this component focuses on understanding the curricular differences between the population for which the test was originally designed and the population to which the test will be exported. As stated in the warrant in Figure 6, one must ask whether the set of tasks on the exported assessment may be viewed as a random sample of the content from the universe of the domain. In other words, can parallel forms of the assessment be viewed as equally representative of the assessed domain? In so doing, can test users have confidence that the interpretations made from the scores in fact make a statement about test takers' knowledge about the assessed domain?

Further, in extrapolation, one asks whether the users of the scores can be assured that the test is not speeded for the test takers taking the exported assessment. Here, *speededness* refers to a situation where the time limits in the test do not allow a large group of test takers to carefully interact with all the items presented in the test. In such situations, the test might be partially measuring the speed of test taking rather than ability on the assessed construct. Variation in test scores reflect real differences in the examinees' knowledge, skills, and abilities related to the construct

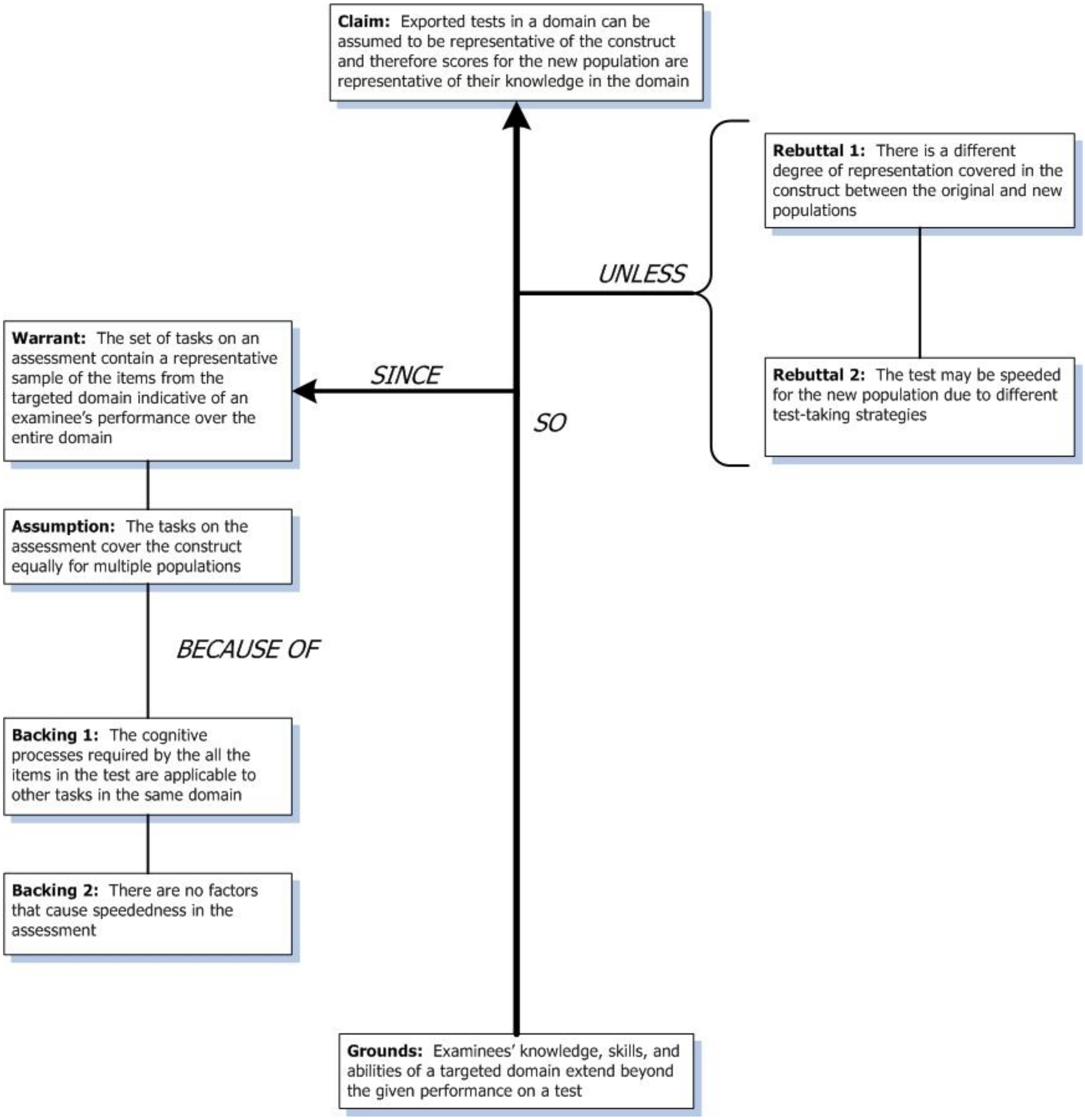

Claim: Exported tests in a domain can be assumed to be representative of the construct and therefore scores for the new population are representative of their knowledge in the domain
UNLESS
Rebuttal 1: There is a different degree of representation covered in the construct between the original and new populations
Rebuttal 2: The test may be speeded for the new population due to different test-taking strategies
SINCE
SO
Warrant: The set of tasks on an assessment contain a representative sample of the items from the targeted domain indicative of an examinee's performance over the entire domain
Assumption: The tasks on the assessment cover the construct equally for multiple populations
BECAUSE OF
Backing 1: The cognitive processes required by the all the items in the test are applicable to other tasks in the same domain
Backing 2: There are no factors that cause speededness in the assessment
Grounds: Examinees' knowledge, skills, and abilities of a targeted domain extend beyond the given performance on a test

Figure 6. Toulmin diagram for extrapolation.

Component 6: Utilization

In a validity argument framework, *utilization* refers to how the scores from a test will be used. Bachman (2005, pp. 18-19) reflected on Kane's framework and argued that there are four warrants behind utilization: (a) the score interpretations are relevant to the decisions that are to be made such that the KSAs that are assessed are directly relevant to the constructs measured in the targeted language of the assessment. (b) the score interpretations are backed by the highest probability that the correct decisions are made by the assessment; (c) beneficial outcomes are the appropriate consequence of the use of a particular assessment, and the beneficiaries are the test takers, the test users, and society at large; and (d) sufficient information about the test takers' skill in the construct(s) of interest come from the assessment so that the users of the assessment can make appropriate outcome decisions. Utilization has direct implications for the new populations. Unless the evidence necessary to support exporting the assessment to a new population is articulated in a very specific way, ambiguities in the IUA will occur, which may raise questions of validity regarding the new population to which the assessment will be given as well as for the new users of the scores.

Figure 7 illustrates the Toulmin arguments for utilization. The warrant suggests that scores may yield valid score-based inferences for the multiple populations. The backing statements specify that the IUA is explicitly stated, that is, scores rendered by the assessment are used for the intended purpose of the test for the multiple populations to which it is administered. The items have been reviewed to ensure that they are appropriate for the new population and comparability studies have been conducted to ensure that the inferences drawn from those scores are comparable for all populations to which the assessment is administered.

As an example, in a licensure situation, those seeking to become certified public accountants (CPA) in the United States must pass the Uniform Certified Public Accountant Examination. This type of certification allows individuals to work in a variety of areas of finance including financial planning, financial accounting, and tax preparation, to name a few. However, the utility of this test would be called into question if exported to the United Kingdom or France, where different rules and regulations may exist for similar professions.

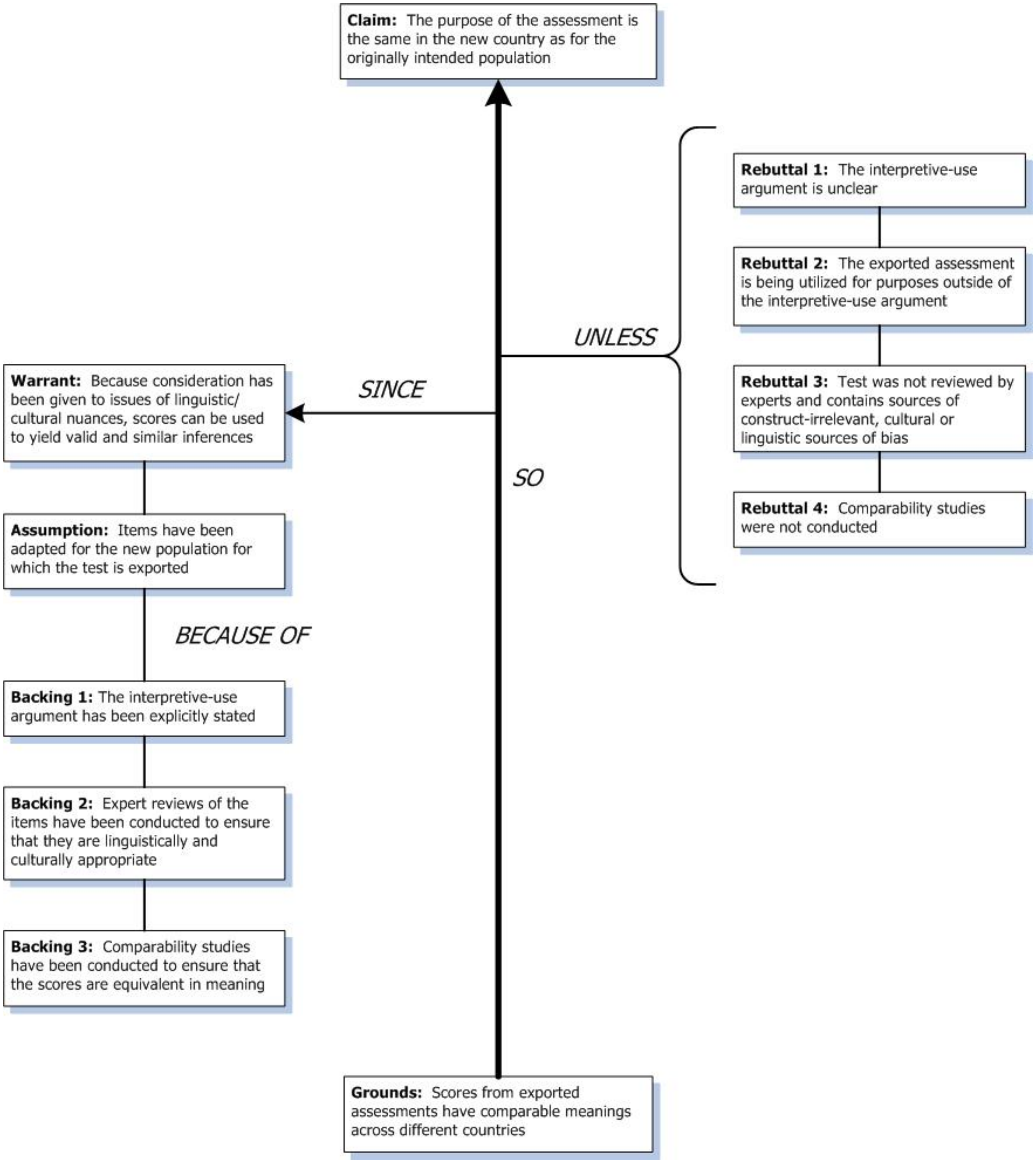

Claim: The purpose of the assessment is the same in the new country as for the originally intended population
UNLESS
Rebuttal 1: The interpretive-use argument is unclear
Rebuttal 2: The exported assessment is being utilized for purposes outside of the interpretive-use argument
Rebuttal 3: Test was not reviewed by experts and contains sources of construct-irrelevant, cultural or linguistic sources of bias
Rebuttal 4: Comparability studies were not conducted
SINCE
Warrant: Because consideration has been given to issues of linguistic/ cultural nuances, scores can be used to yield valid and similar inferences
SO
Assumption: Items have been adapted for the new population for which the test is exported
BECAUSE OF
Backing 1: The interpretive-use argument has been explicitly stated
Backing 2: Expert reviews of the items have been conducted to ensure that they are linguistically and culturally appropriate
Backing 3: Comparability studies have been conducted to ensure that the scores are equivalent in meaning
Grounds: Scores from exported assessments have comparable meanings across different countries

Figure 7. Toulmin diagram for utilization.

Conclusion

We designed this framework to address potential threats to validity in the use of exported assessments. We outlined key considerations that require attention when using assessments with multiple populations and illustrated the importance of each component through the use of examples. Given the crucial nature of high-stakes decisions based upon test scores from high-stakes assessments, we suggest following very comprehensive validation procedures to ensure that the inferences made for the new populations taking these assessments have consistent meaning with the original populations. These validation procedures include ensuring that the content domain has been reviewed by experts and that field tests and psychometric investigations (including DIF and measurement invariance analyses) have been conducted to make certain that parallel test forms yield comparable scores. We also suggest empirically evaluating the assumptions underlying psychometric models that are used during the analyses of items (such as the assumption that all test takers within the population use the same combination of KSAs when interacting with the test items) as these assumptions may be violated when assessments are administered to multiple populations (Ercikan & Oliveri, 2013; Oliveri, Ercikan, & Zumbo, 2014). In addition to these validation procedures, we recommend developing test preparation materials that contain clear examples of all item types to enable equal access for members of all potential test-taker populations to help them familiarize themselves with the item types and content contained within the assessment prior to taking the assessment.

We recognize that there are limitations in the ability to satisfy all of the constraints presented in the six components of the framework for tests that may have been exported long before the development of these guidelines, as changes to the assessment may need to be made in a gradual fashion. Another limitation may be associated with the cost of hiring panels of experts or test developers to reanalyze the domain that is to be assessed or create/modify items after pilot or field testing. Other limitations include the inability of recruiting adequate numbers of students to field test items effectively with the intended population so as to uncover construct-irrelevant factors related to cultural or linguistic differences. However, field testing is important as its results can also be used to confirm whether or not a test is unduly speeded for one population or whether a new population has an apparent difficulty in understanding some of the item types that compose the test. Not field testing the multiple populations can lead to higher level challenges related to scores and their interpretations. These challenges include comparability of scores between the original and new populations and whether the scores appropriately represent members of the new population's knowledge of the intended construct.

It is reasonable to expect that there will be cultural differences between the original and new populations, and in terms of assessment, one size will not fit all. A small sample of cognitive interviews may reveal problematic item types for the new population due to cultural or pedagogical differences. We suggest evaluating these potential differences in support of the IUA.

Our framework was inspired by ITC guidelines (ITC, 2005, 2013) and the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014). In the spirit of those guidelines, we suggest that test developers play a key role in reviewing assessments that are to be considered for exportation because they possess the fundamental knowledge of the construct(s) being tested. Best practice indicates that when their efforts are coordinated with experts familiar with the targeted culture for which a test might be exported, items that may be culturally or linguistically loaded can be identified prior to the administration of the assessment. We also recommend collaborating with psychometricians prior to the actual test administration to interpret differences in scores, levels of invariance, and score scales between populations, which may become apparent during field testing. Finally, researchers should be an integral part of the team to corroborate the findings of pretesting so that differences between populations can be added to the accumulation of evidence used to support the fair use of any exported assessment to yield valid score-based inferences.

References

Altbach, P. G., Reisberg, L., & Rumbley, L. E. (2009). *Trends in global higher education: Tracking an academic revolution. Report prepared for the UNESCO 2009 World Conference on Higher Education.* Paris, France: UNESCO.

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.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly, 2*(1), 1–34. doi: 10.1207/s15434311laq0201_1

Bracken, B. A., & Barona, A. (1991). State of the art procedures for translating, validating, and using psychoeducational tests in cross-cultural assessment. *School Psychology International, 12,* 119–132.

Brislin, R. W. (1980). Translation and content analysis of oral and written materials. In H. C. Triandis & J. W. Berry (Eds.), *Handbook of cross-cultural psychology* (pp. 389–444). Boston, MA: Allyn & Bacon.

Brislin, R. W. (1986). The wording and translation of research instruments. In W. J. Lonner & J. W. Berry (Eds.), *Field methods in cross-cultural research* (pp. 137–164). Thousand Oaks, CA: Sage.

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*(1) 3–13.

Clauser, B. E. (2000). Recurrent issues and recent advances in scoring performance assessments. *Applied Psychological Measurement, 24,* 310–324.

Educational Testing Service. (2009). *ETS international principles for fairness review of assessments: A manual for developing locally appropriate fairness review guidelines in various countries.* Princeton, NJ: Author.

Educational Testing Service. (2013). *Boletín de Información e Instrucciones del examen EXADEP.* Princeton, NJ: Author.

Educational Testing Service. (2015). *ETS standards for quality and fairness.* 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*(2), 24-35.

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 and test use: An international dialogue on educational assessment, accountability, and equity* (pp. 69–86). Bingley, UK: Emerald Publishing.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6,* 304–312.

Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist, 52,* 1115–1124.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment, 10,* 229–244.

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.

Helms, J. E. (1997). The triple quandary of race, culture, and social class in standardized cognitive ability testing. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment* (pp. 517–532). New York, NY: Guilford Press.

International Test Commission. (2005). *International Test Commission guidelines for translating and adapting tests.* Retrieved from http://www.intestcom.org/files/guideline_test_adaptation.pdf

International Test Commission. (2013). *ITC guidelines on test use.* Retrieved from http://www.intestcom.org/files/guideline_test_use.pdf

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50,* 1–73.

Laing, S. P.,& Kami, A. (2003). Alternative assessment of language and literacy in culturally and linguistically diverse populations. *Language, Speech, and Hearing Services in Schools, 34,* 44–55.

Mislevy, R. J. (2010). Some implications of expertise research for educational assessment. *Research Papers in Education, 25*(3), 253–270.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice, 25*(4), 6–20. doi: 10.1111/j.1745-3992.2006.00075.x

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (1999). *On the roles of task model variables in assessment design*. Los Angeles, CA: National Center for Research on Evaluation , Standards, and Student Testing .

Oliveri, M. E., Ercikan, K., & Zumbo, B. D. (2014). Effects of population heterogeneity on accuracy of DIF detection. *Applied Measurement in Education. 27,* 286-300.

Oliveri, M. E., Gundersen-Bryden, B., & Ercikan, K. (2012). Scoring issues in large-scale assessments. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large-scale assessment in education: Theory, issues and practice* (pp. 143–153). New York, NY: Routledge/Taylor & Francis.

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.

Rhodes, R. L., Ochoa, S. H., & Ortiz, S. O. (2005). *Assessing culturally and linguistically diverse students: A practial guide.* New York, NY: Guilford Press.

Toulmin, S. (1958). *The uses of argument.* Cambridge, UK: Cambridge University Press.

van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. *European Psychologist, 1,* 89–99.

van de Vijver, F. J. R., & Leung, K. (1997a). *Methods and data analysis for cross-cultural research.* Thousand Oaks, CA: Sage.

van de Vijver, F. J. R., & Leung, K. (1997b). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (Vol. 1, 2nd ed., pp. 257–300). Boston, MA: Allyn & Bacon.

van de Vijver, F. J. R., & Tanzer, N. K. (1997). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 47,* 261–329.

von Davier, A. A., & Oliveri, M. E. (2013, April). *Psychometrics in support of a valid assessment of linguistic subgroups: Implications for the test and sampling designs.* Paper presented at the meeting of the National Council on Measurement in Education, San Francisco, CA.

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*(2), 147–170. doi: 10.1177/0265532209349465