2016

Cultural-Linguistic Test Adaptations: Guidelines for Selection, Alteration, Use, and Review

S. Kathleen Krach, Michael P. McCreery and Jessika Guerard

Cultural-Linguistic Test Adaptations:

Guidelines for Selection, Alteration, Use, and Review

S. Kathleen Krach, Ph.D., NCSP

Florida State University

skrach@fsu.edu

1114 West Call Street, Tallahassee, FL 32304

Michael P. McCreery, Ph.D.

University of Nevada Las Vegas

michael.mcreery@unlv.edu

368 CEB, 4505 S. Maryland Parkway, Las Vegas, NV 89154-3001

Jessika Guerard, B.A.

Florida State University

jg15j@my.fsu.edu

1114 West Call Street, Tallahassee, FL 32304

Keywords: Tests, Translations, Psychometrics, Multilingual, Multicultural

Correspondence: All correspondence should be addressed to: S. Kathleen Krach, Ph.D., NCSP

Word Count: 8,000

Authors note: The authors would like to thank the reviewers for their input in this article. Specifically, the recommendations by Bruce Bracken were valued.

Abstract

In 1991, Bracken and Barona wrote an article for School Psychology International focusing on state of the art procedures for translating and using tests across multiple languages. Considerable progress has been achieved in this area over the 25 years between that publication and today. This article seeks to provide a more current set of suggestions for altering tests originally developed for other cultures and / or languages. Beyond merely describing procedures for linguistic translations, the authors provide suggestions on how to alter, use, and review tests as part of a cultural-linguistic adaptation process. These suggestions are described in a step-by-step manner that is usable both by test adapters and by consumers of adapted tests.

Cultural-Linguistic Test Adaptations:
Guidelines for Selection, Alteration, Use, and Review

HWÆT, WE GAR-DEna in geardagum,

þeodcyninga þrym gefrunon,

hu ða æþelingas ellen fremedon!

(Beowulf, in Old English, Klaeber, 1922)

LO, praise of the prowess of people-kings

of spear-armed Danes, in days long sped,

we have heard, and what honor the athelings won!

(Beowulf, in modern English, Gummere, 1910)

Although both of these samples are in English, one may be unreadable to modern-day speakers of the language. That is because only one word ("in") from the old-English version is still the same in the modern-day version. So, why are the two versions (both in English) so different? These changes did not happen quickly, but came about steadily over time due to several factors. One is that new words are always coming into use to serve previously unknown purposes. For example, the 2015 Oxford Dictionary's English word of the year was "emoji," a word that barely existed in the lexicon 10 years ago, and not at all 50 years ago. In addition, as new words enter a language, other words exit due to shifts in popularity (e.g., words like "boffin" or "bouffant"). Words may be lost because they no longer serve a current purpose (e.g., "45 rpm adapter" for converting a record player). Thus, it is clear that a single language can change

dramatically from decade to decade, eventually resulting in such changes as seen in the examples above.

These types of changes happen not only to the English language, but to other languages as well. As each language can be presumed to change at a similarly rapid rate, then attempting to translate between two (or more) languages at a time may prove to be a difficult challenge. Add to this the possibility of needing translated versions for more than 7,000 independent languages worldwide (Ethnologue, 2008; http://www.ethnologue.com/). And these are only simple linguistic concerns. Sometimes the literal meaning of a word (denotation) and the value accompanying the word (connotation) may change within a single language, as well as from language to language. One example of an English word that has different connotations depending on context would be "feminist." This is because the value attached to this word by some individuals is positive, while others view it as negative.

Such language issues are complicated enough when trying to translate a work of fiction or technical documentation, but they become compounded with test and measurement issues when trying to make a test developed for a specific cultural-linguistic group available for a different language and / or culture. One such issue is the translation of content; one may translate an item from its origin language and maintain the item's cultural value, but may lose the value (connotation) of or level (difficulty) of the construct of interest being measured (Hambleton, 2005; Sattler, Oades-Sese, Kitzie, & Krach, 2014). For example, if the goal is to measure reading ability, then translated items should have the same number of phonemes. If they do not (e.g., as happens when changing the one-syllable English word "car" to the two-syllable Spanish word "coche,") then this changes the item's difficulty level.. This same translation may also subtly change the value of the task as well. For example, the use of the Castillan-derived word "coche"

may be seen as more “authentic” by native Spanish speakers than the use of the English-influenced word “carro” or the Germanic-influenced version “auto” (Roggen, 2014).

Value differences may also make it difficult to translate the actual meaning of certain words. For example, translating a rating scale item of “I feel blue” from English does not allow for a direct word-for-word translation into Chinese. Instead, the item has to be translated as “I feel sad,” which is similar to the same construct but does not hold the same value (Ren, Amick, Zhou, & Gandek, 1998). And, finally, certain activities are valued differently for different groups. For example, a rating-scale item stating, “My child prepares a simple meal’” may produce dramatically different answers depending on the cultural group to whom it is administered. This is because, in some countries, children start preparing food at an earlier age than others, thereby changing the age-expectation value of the task. In other countries, only women prepare the meals, thereby changing the gender-expectation of the task. Finally, the idea that preparing food is a solitary experience (and not a performed as a group or family) can also be culturally loaded.

Given all of these considerations, the term “test translation” is too narrow a description for the modern process. Instead, the term “test adaptation” is more appropriate. Test adaptation includes all considerations of the cultural-linguistic transfer of a test from one group to another instead of the singular aspect of simple translation (Hambleton, 2005). The purpose of this article is not to dissuade the reader from using tests designed for other populations. Instead, it is to provide guidance in the daunting task of making decisions regarding test adaptation. Thus, this paper attempts to meet two goals. The first is to serve as a primer for test adapters to consider when using a test originally designed for a different culture / language. The second is to provide test users assistance in selecting, using, and interpreting adapted tests when necessary.

Step One: Familiarizing Yourself with Recommendations from the Field

Bracken and Barona (1991) published one of the first multi-step guides for properly translating tests. One difference between their work and the current paper was that the procedures they recommended were for translating tests largely for intranational use, not international. Specifically, they discussed translating tests to be used within the United States with speakers of languages other than English. This population is different from an international audience in that, although hundreds of languages are spoken in the United States, the assumed goal of those who speak these languages is to eventually speak English. This is not true of speakers of languages other than English who live in their native countries. This difference is subtle but important. The most important aspect of this difference is not linguistic but cultural. Specifically, when test items are written and published for use in the United States, these items now addresses the extent to which the individuals tested conform to U.S. cultural norms. For these reasons, their work focused on the linguistic translation and not the cultural-linguistic adaptation that encompasses the current work.

In Bracken and Barona (1991), their first step was to enlist a bilingual individual familiar with the test to conduct a word-for-word (or meaning-for-meaning) translation from the origin test to the target test. The second step consisted of having someone who had never seen the origin test perform a back-translation from the target version to the origin language. These steps were repeated as often as necessary until only a minimal gap between the origin test and its back-translation existed. The third step was to have a multinational or multiregional bilingual committee review both versions to ensure the target test met the requirements for content and construct similarity (Bracken & Barona, 1991).

Upon approval of the target test by the committee, the next step was to conduct pilot testing in which the target instrument was administered to individuals who shared the same language but hailed from different cultural, social, and economic backgrounds. Upon completion of pilot testing and related changes, field-testing was completed to evaluate psychometric issues such as reliability, validity, and normative data collection (Bracken & Barona, 1991). The last step focused on test version equivalency using statistical techniques such as factor analyses and multi-trait-multi-method analyses (Bracken & Barona, 1991).

Following the work of Bracken and Barona (1991), other authors (Butcher, 1996; Geisinger, 1994) built similar sets of guidelines for test translation. Figure 1 provides a unified combination of these steps. Following their combined work, many professional organizations began to establish definitive rules and ethical restrictions for test adaptation. For example, in 1999, the International Test Commission (ITC) developed its testing regulations, as did a joint partnership of the American Educational Research Association (AERA), American Psychological Association, and National Council on Measurement in Education. In 2001, the U. S. Office of Minority Health also added standards for providing competent services to culturally and linguistically diverse clients. More recently, the ITC (2005) updated its standards, as did the National Association of School Psychologists (NASP, 2010) and the International School Psychology Association (ISPA, 2011). Table 1 provides a list of the guidelines from each of these organizations specific to test interpretation. Please note that these guidelines focus mostly on the following concepts: 1) who can translate and administer the tests (e.g., fluency, testing knowledge, cultural competency and 2) what type of tests to use (e.g., culturally appropriate, psychometrically sound, etc.). There is an additional emphasis that documenting the methods used for translating an instrument should be evidence-based as well as culturally and

linguistically appropriate. Although these guidelines are helpful, these organizations provide no specific information on what these evidence-based practices are, nor what constitutes cultural and linguistically appropriate approaches.

Gudmundsson (2009) provides one of the most recent and comprehensive sets of guidelines for adapting tests. He laid out an eight-step procedure for both test translation and adaptation. The first step was selecting an instrument for translation and adaptation, with an emphasis on instruments that are psychometrically sound. The second step emphasizes the need to select skilled translators who are fluent in both the language of, and knowledge about, the origin test. The third step requires the selection of subject-area experts to ensure that the construct remains sound across both versions. The fourth step emphasizes the need to be thoughtful in the method of adaptation (word-for-word, meaning-for-meaning, construct-for-construct), with a goal of decreasing overall bias in the target test. The fifth step calls for the use of the selected method of adaptation; the sixth focuses on reducing bias in the target version. The seventh step includes a pilot study with a focus on item analysis and administration techniques. The final step emphasizes psychometrics such as reliability, validity, and equivalency. Although Gudmundsson's (2009) work is well-considered and follows the ethical and legal structures set forth by the professional organizations, it is non-specific. The methods described in the rest of this article are designed to provide more specifics on the process by integrating previous research as well as ethical best practices.

Step Two: Determining rationale for testing

Although the first step in other works is test selection, (Bracken & Barona, 1991; Butcher, 1996; Geisinger, 1994), it is not even the second step in the current model, which focuses more on determining why and what the practitioner would like to test rather than

focusing on a specific instrument. This model uses the work of Kumman (2005), who developed a Test Context Framework (Figure 2) as the basis for all test adaptation decisions. According to this framework, every test is embedded in a system comprising laws and ethics governing its use; political and economic reasons for its development; technological constraints and affordances; and educational, social, and cultural constructs that guide its purpose. Without considering the reason for adapting a test using a framework like Kumman's (2005), all other decisions are stalled.

The following demonstrates how a single test can be embedded across several aspects of this framework. In this example, an achievement test is administered to a child in primary school. Ethics and laws govern the manner in which the test is administered. The specific items on the test are determined by politicians' decisions on the academic standards for that child's level. The child's performance may influence how much money the school receives. The test may be administered by a computer, but only if the school has the infrastructure to support such technology.

Finally, there may be test-taker issues that influence test administration and interpretation (Sattler et al., 2014). Amongst these considerations is the possibility that the child may struggle with the test because she is from a cultural group where academic support is not offered in the home. On the other hand, she may have social pressure to be successful on the test, so many supports are provided by the school and family (van de Vijver & Poortinga, 1997). Also of note is the very act of having a school psychologist administer a test may create cross-cultural barriers. For example, if children are raised not to look adults or members of the opposite sex in the eye, that may influence test performance. In addition, test-taker interactions with specific test tasks may be quite different from person to person. A specific example would be adapting a

timed-test for use within a culture that emphasizes accuracy over speed, thereby decreasing the validity of the instrument.

There are some simple guidelines for determining the need to create a new test adaptation. First, it is important to make sure that the publisher does not already have a translated version available. This may seem intuitive, but finding a published, adapted test,is not as easy as entering the name of a test into a search engine and getting a list of all versions available. Search engines modify searches based on the country you are searching from, so that in the U.S., your search engine might default to "publisher.com." Many publishers do not list all of their possible translations on their "publisher.com" website. Instead, in the U.S., the publisher may only list the U.S. English and Spanish versions. Thus, if you want a version for a different language, you must navigate to that language's native country's version of the website; for example, if there is an Australian version of the test, it might be on listed on "publisher.com.au."

By surveying publishers directly, Krach, Doss, and McCreery (2015) found that, of 45 social / emotional / behavioral tests examined, there were more than 143 published versions available. While a few tests only had an English version, some had as many as 30 different translated or adapted versions available. It should be noted that, although some of the different published versions of the test were full adaptations, many were only simple translations. For example, in the Krach and colleagues (2015) study, 26 of the 42 tests had versions in languages other than English, but these provided only a translation of the test with no new normative data. Thus, just because a publisher provides a translated version does not mean that the test was adapted. In such cases, it may be necessary to use information from both Figure 1 and Table 1 to review what has been done before determining whether the test adaptation is sufficient for your needs.

If a publisher's translated / adapted version of a test is not sufficient for your needs, please note that better instruments for your target population than the one you have chosen as your origin test may be available. Do not ignore tests that were originally developed for the target language or cultural group of interest. Tests developed in your client's country of origin should be seriously considered before trying to adapt a test on your own. However, though published tests exist for many cultural-linguistic groups, the precise one you need may not yet have been adapted or even written, given the 7000 languages spoken on this planet. It is not financially practical for publishers to undergo all of the formal steps of a proper cultural-linguistic test adaptation for all language options. If this is the case for you, then you might choose to adapt a test on your own; however, it should be noted that adapting a test should be your final option.

If you must adapt a test on your own, all components of Kumman's (2005) framework must be considered. Out of these, the two of utmost importance are: 1) for what reason is a specific test selected for adaptation, and 2) in what ways does the reason influence the adaptation process. For example, if a test is selected for "high stakes" reasons (e.g., holding a child back a level, teacher pay, making a diagnosis, etc.), then any test adaptation should undergo the most rigorous requirements possible in the adaptation process. This more stringent approach to adaptation would include teams of experts working on the translation, determining equivalency between the origin and target versions and collecting normative data for comparisons. However, if the stakes are lower (e.g., monitoring progress, screening, etc.), then a less formal adaptation process involving translation, collecting data, and determining equivalency may be acceptable. However, the amount of data needed and the methods used to determine equivalency might be less burdensome.

In the next two steps, the article will discuss the recommended professional practice in test adaptation, as well as ideal (and less than ideal) adaptation procedures.

Step Three: Appropriate Adaptation Procedures

This adaptation procedures section is divided into two sets of guidelines for adapting tests. The first includes a list of ideal guidelines and the second includes a list of less than ideal guidelines. Please note, even if you choose to use the less than ideal guidelines as your standard, you should still strive to reach the goals listed in the ideal version.

Ideal adaptation procedures. There is no mystery surrounding how to culturally and linguistically adapt tests. The list provided in Figure 1 is an excellent place to start. The following describes each of these steps in more detail, adding information that has been updated since the authors cited in Table 1 first developed their plans.

Choose origin test. According to the first step, it is vital to choose the appropriate origin test from the beginning. Arguably, the best way to do this would be to build all targeted cultural-linguistic variations of an origin test concurrently with the development of the original instrument (Solano-Flores, Trumbull, & Nelson-Barber, 2002). This would ensure that all cultural and linguistic issues are addressed prior to development, and no version would be a lesser version. All items included originally would have a reduction in bias; therefore, no additions, subtractions, or item alterations would be needed to ensure cultural equivalency. It should be noted that no test will be completely free of bias; instead, the goal should be to reduce bias as much as possible.

When concurrent test development is not feasible, then the origin test must be chosen carefully to ensure that it can be altered to measure a similar construct in the target version; additional considerations include the need to ensure that the origin test is psychometrically sound

and uses simple items with basic instructions (Bracken & Barona, 1991). Sometimes, the selection of the origin test is limited by variables such as 1) the dominance of the test (e.g., the Wechsler scales may be chosen because of popularity), 2) the type of data required (e.g., special education law or the diagnostic manual), or 3) a lack of other assessments that measure the same construct. When the origin test options are narrowed, then it is vitally important that all remaining adaptation steps are followed with fidelity.

Translate the test. Next, the process of starting the initial translation begins. It is at this earliest stage that the focus should be on bias (Brislin, 1980). Each of the three types of bias comes from a different source; Van de Vijver and Poortinga (1997) provide a helpful table that is reproduced in Table 2.

At the simplest level, one may assume that there is limited bias in the construct measured (e.g., belief that phonemic awareness is a universal construct) and / or assume there is limited bias in the method of assessment (e.g., assume rating scales are a universal method of collecting data). However, such assumptions must always be confirmed empirically. If they are not, then results from the instrument cannot be interpreted with fidelity. In the phonemic awareness example used here, the assumption of a lack of construct bias would be erroneous because not all languages use the same phonemes. This is true for a method bias assumption when using rating scales; not all cultural groups respond the same way (Hui & Triandis, 1989).

Once it is determined that neither construct nor method bias is an issue, the test adapter should focus on item bias (Van de Vijver & Leung, 1996). Item bias occurs when there is either a poor translation of the item or when the item may have different cultural–specific interpretations. However, empirical investigations have shown that when test developers focus on both construct

and how that construct manifests across cultural groups, they are able to minimize item bias (Byrne & Van de Vijver, 2010).

The goal in the earliest steps is not to determine bias but to prevent it. In the later steps of review and validation, the researchers will run analyses to examine each source of bias. In this earliest step, the test adapter is simply determining whether it is even feasible to make a version of an instrument for a new cultural-linguistic group, or if the test will be too biased for an adaptation to be possible. When it appears that bias will keep a literal translation from being possible, there are only two choices: adapt only parts of the instrument, or create an entirely new instrument (Van de Vijver & Hambleton, 1996).

Review. If either a part or the whole of a test can be adapted without bias, then one may move forward with a cultural-linguistic translation. After that, it is vital that the adapted instrument be reviewed prior to moving forward with any additional steps. One common method of review is through a process called back-translation. Brislin (1980) describes translation as a process of taking a document translated from one language (L1) to another language (L2), and back translation as the process of translating it back again (L2 to L1). In addition to back-translation, the review step provides the opportunity for a qualitative analysis of the test items. A panel of experts should provide a comprehensive review of the translated items that are problematic in terms of either cultural or linguistic equivalency (Geisinger, 1994). Members of the panel should review the information separately and come together later to make joint recommendations.

Alterations. The next step is to make appropriate alterations to the translated test as needed. The test developer starts by making changes based on the panel's recommendations. The revised version of the materials then goes back through the panel review process repeatedly until

all concerns are addressed. After the panel clears the work, pilot testing is needed. The goal of pilot testing is to understand how the translated test works in a real-world setting. Van Teijlingen and Hundley (2002) provide a good outline on the steps of pilot testing, writing that pilot versions of a test should be administered in the same manner as the final version. However, afterwards, test-takers should provide feedback on item difficulty, clarity, and confusion, as well as answer questions about the testing process and expectations. Following the collection of pilot data, the test adapters should make changes that reflect the feedback from the pilot tests and address any bias or psychometric concerns that arise from the collected data.

There are several possible techniques for assessing the pilot data for bias. One method to determine both types of equivalence across the construct and the method of testing; these are referred to as structural equivalence and measurement equivalence (Byrne & van De Vijver, 2010). To have structural equivalence, the meaning of the construct measured must be independent from both cultures (Van de Vijver & Tanzer, 2004), and all of the construct's facets, dimensions, or underlying structures are presumed to be equal across cultures (Byrne & van De Vijver, 2010). Structural equivalence is mostly a theoretical concept that is difficult to quantify.

So, despite the importance of structural equivalence, overall test equivalence is traditionally determined through measurement equivalence, which provides an empirical method to analyze construct structures across cultures (Byrne & van De Vijver, 2010). Specifically, measurement equivalence is the evaluation of factorial structure and loading (e.g., do items that load on one factor for the English version still load on that factor for the Spanish version) as well as item content and mean equivalency (i.e., do the means and standard deviations for each item equate across cultural-linguistic groups, or does one group score higher or lower). For more information on the specific statistical procedures for each of these techniques, see Kankaraš, &

Moors (2010). It is important to note that measurement equivalence implies that different cultures are measuring the same construct. In other words, one may achieve structural equivalence but not measurement equivalence, but one must achieve structural equivalence to achieve measurement equivalence.

Validation and Publication. It is the responsibility of the developer to provide documentation of all steps taken to ensure the instrument is psychometrically sound. As part of this, data from any pilot, bias, or validation studies should be provided so that users can interpret the appropriate use of the instrument for themselves (AERA et al., 1999). In the case of cultural-linguistic test adaptation, it is the developers' responsibility to also disclose the translation and review procedures in addition to any validation or equivalency studies conducted as part of the adaptation process (Geisinger, 1994). Specifically, test publishers must disclose attempts to overcome the different types of bias embedded in the adaptation process (see Table 2 for specific sources of bias). A reviewer would peruse all of the potential types of bias described in Table 2 and expect to find corresponding information in the manual to document how the test developers considered, overcame, or explained a work-around for this type of bias.

Administration. The final step is to oversee initial administrations of the test. This means that the test adapter is responsible for determining the level of training needed to administer the test, setting up opportunities for training, and providing a method for users in the field to offer feedback on the test after publication (Geisinger, 1994). In addition, the adapter is responsible for disseminating updates to the adapted tests to users in field as new information becomes available. Finally, the adapter must consider the need to update the test periodically to accommodate problems associated with societal changes (Van der Velde, Feij, & van Emmerik, 1998) and the Flynn (1988) effect.

Less than ideal adaptation procedures. The previous section on the ideal methods for adapting a test works well for professional publishers who have the time and resources to complete all of the steps. However, individuals who need a translation of a test that is unavailable or nonexistent may fall back upon a less than ideal method of test adaptation. Just as with the ideal method of adapting an instrument, Figure 1 should be your guide.

Choose origin test. As before, the process starts by choosing the test you wish to adapt. This step is virtually unchanged from the ideal version. However, you do not have the freedom of choice that the origin test publisher may have unless the copyright of the origin test allows for adaptation. If the test copyright does not allow for adaptation, you must seek permission from the publisher before starting the process.

Translate the test and review. Once permission is granted, it is time to start the translation itself. The translation (and back-translation) of the instrument remains the same (Brislin, 1980), as does the need to consider bias (construct, method, and item). Although having a comprehensive panel to help with this process may be impossible and impractical, it is recommended that the translation / back-translation be conducted by two different people. You should have a third person involved with the review process. Each of these individuals should be knowledgeable about both languages and cultures (APA, 2010; NASP, 2010). The final reviewer should work independently from anyone else involved in this process.

Alterations and validation. Alterations should be made based upon the final reviewer's comments. This may include adding, removing, or substituting items. Validation procedures for the less than ideal adaptation are the biggest area of difference between the two methods. For instruments adapted using this method, normative data from the original version can no longer be used (Bracken & Barona, 1991; Rhodes, Ochoa, & Ortiz, 2005). Because there is no need to

worry about the normative effect, changing items is not problematic. However, if the test user is going to use a raw score-to-raw score comparison, the user must either administer all items (no basals / ceilings) or make any changes to both versions to ensure accuracy. For example, in the "car / coche" translation described above, the problem is that "car" has three phonemes and "coche" has four. The adapter's choice is to do a literal translation, finding one word in each language that has three phonemes, or to change "coche" to a concrete common noun that has three phonemes.

As part of the validation process, the test adapter will still want to create a pilot version of this test. Preferably, they would do this with at least five test takers who are fluent in both languages. The pilot group should not include any referred individuals who need the test, because the test has not been validated and should not be used for diagnostic purposes. After the pilot test, changes should be made as needed and reliability data should be analyzed.

Publication and administration. The adapted version of an instrument cannot be distributed without the publisher's permission. Even though the test adapter will not be selling or distributing the test, the individual should write down all of the steps used in the adaptation process. This information is needed for future users, as well as for the adapter's own use, in case there is a need to defend the findings from the adapted test in a court of law. There is another copyright consideration in the use of an adapted version of a test. The user may still be responsible for purchasing the blank record form of the origin version to accompany the target version; the two should be stored together. This is because the right to use still belongs to the original publisher. Finally, the test adapter must ensure that the person administering the test is both fluent in the target language and knowledgeable about testing procedures (Hallberg, 1996; Rhodes et al., 2005).

Even less than ideal version. There is no "even less than ideal version" that you can use in this process. For example, you should not be conducting "ad-hoc" translations as part of your practice. Ad-hoc testing does not follow any of the guidelines listed in Figure 1, nor does it address any of the sources of bias found in Table 2. Thus, any test findings cannot be considered psychometrically or theoretically sound, and are not fit for use. Although the desire to use a short cut may be strong, it is strenuously not recommended.

Unfortunately, the use of ad-hoc translations is a common procedure among school psychologists in the U. S. (about 50% say they have done it; Ochoa, Riccio, Jimenez, de Alba, & Sines, 2004). Compounding the problem is that most school psychologists have not been trained to identify and recruit appropriate translators (Ochoa, Gonzalez, Galarza, & Guillemard, 1996). Inappropriate translators include individuals such as secretaries and janitors (Paone, Malott, & Maddux, 2010), the referred child or older sibling (García-Sánchez, Orellana, & Hopkins, 2011; Tse, 1995), friends of the child or family (Lynch & Hanson, 1996), or foreign-language teachers at the school (Swender, 2003). Appropriate translators should be individuals who are trained for this specific task and are thoroughly aware of the national standard for translator code of ethics (e.g., American Translators Association Code of Ethics (n.d.). These individuals should be certified and / or licensed in the field of translation services (if applicable in your area).

Step four: Drawing Conclusions from Adapted Test Data

Once a test has been adapted, an interpretation plan must be established. When interpreting results from any tests, it is vital to consider corroborating data in decision-making (NASP, 2010). For example, if a teacher reports that a child is performing well in school, but a test score says otherwise, then it is vital to investigate further to determine why there is a discrepancy between the two sources (Sattler, 2008). This is, and should be, common practice.

Please note that in this practice, each piece of data may not hold as much weight as every other piece in the decision-making process. For example, scores from a psychometrically sound instrument might be weighed more heavily than interview data from a novice teacher; however, interview data from an experienced teacher should be more strongly considered than data from a poorly adapted test.

The most important consideration for interpreting a culturally-linguistically adapted test is ensuring that you get the weightings correct for a given piece of data. These weightings are based purely on clinical judgment, with a primary concern that the further away from the "ideal" adaptation of the test, the less the findings from that instrument should be weighted in decision-making processes (van de Vijver & Poortinga, 1997). Instead, when the origin and target versions of a test differ considerably, then users should depend more on outside data sources for corroborative support (e.g., review of records, interviews, observations, etc.). In addition, tests that have been adapted using the "less than ideal method" should be weighted with less consideration than those done using the "ideal method;" while any which use ad-hoc data should be disregarded. Studies show that, for the most part, clinicians weight data differently when making interpretations for cultural-linguistically diverse test-takers. In a study by Sotelo-Dynega and Dixon (2014), 98.3% considered the validity of the test scores in their interpretation. The majority (55.8%) also included informal assessments as supplemental materials for consideration in their analysis. These findings support that test users understand the need for best practice in test interpretation.

Discussion

In 1991, Bracken and Barona wrote an article for School Psychology International focusing on "state of the art" procedures used in translating and using tests for multiple languages. This

seminal piece from 25 years ago led to the creation of a barrage of test translation / test adaptation policies, guidelines, and ethical requirements that have been developed and updated ever since. The current article seeks to build on their original work by providing more specific suggestions and a more current analysis of recent literature. This article attempted to establish empirically based procedures beyond simple linguistic translations with a focus on more complicated cultural-linguistic adaptation. These suggestions have been described in a step-by-step manner for both ideal and less-than-ideal adaptation methods.

The first step in the process was to familiarize oneself with all of the legal and ethical considerations around using tests with multilingual and multicultural populations (AERA, APA, & NCME, 1999; ISPA, 2011; ITC, 2005; NASP, 2010). The next is to determine why a child needs to be tested and what systemic issues should be considered as part of any test selection, adaptation, and interpretation (Kumman, 2005). These considerations help to determine whether an ideal or a less than ideal method of test adaptation is needed. It is only once these two steps are completed that the appropriate adaptation process may begin.

When adapting a test, there are similar procedures to follow for both ideal and less than ideal methods to decrease test bias: 1) Choose origin test, 2) translate the test, 3) review, 4) alterations, 5) validation, and 6) publication (Bracken, & Barona, 1991; Butcher, 1996; Geisinger, 1994). Table 2 describes sources of test bias in more detail, and Figure 1 provides more guidance on each of these procedures. There is no even less than ideal method of test adaptation. Specifically, ad-hoc test translations should never be used for diagnostic purposes (Ochoa et al., 2004; Rhodes et al., 2005).

Finally, when using adapted tests, be careful when drawing conclusions from the findings. Even when adapted using ideal methods, the target version of a test will not be as good as the

original test. Therefore, test users should use professional judgment and make weighted decisions utilizing multiple data sources (Sattler, 2008). Even if the test adaptation process goes perfectly, test users must interpret the data using best practices or the findings will be suspect.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

American Psychological Association (APA, 2010). American Psychological Association's Ethical Principles of Psychologists and Code of Conduct. Washington, DC: Author.

American Translators Association (n.d.). Code of Ethics. Alexandria, VA: Author. Retrieved from https://www.atanet.org/governance/code_of_professional_conduct.php

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. doi: 10.1177/0143034391121010

Brislin, R. W. (1980). Translation and content analysis of oral and written material. In H. C. Triandis & J. W. Berry (Eds.), Handbook of cross-cultural psychology (Vol. 1, pp. 389–444). Boston: Allyn & Bacon.

Butcher, J. N. (1996). Translation and adaptation of the MMPI–2 for international use. In J. N. Butcher (Ed.), International adaptations of the MMPI–2: Research and clinical applications. (pp. 26–43). Minneapolis: University of Minnesota Press

Byrne, B. M., & van De Vijver, F. J. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. International Journal of Testing, 10, 107-132. doi: 10.1080/15305051003637306

Flynn, J. R. (1998). IQ gains over time: Toward finding the causes. In U. Neisser (Ed.), The rising curve. Long term gains in IQ and related measures. (pp. 25–66). American Psychological Association, Washington, DC

García-Sánchez, I. M., Orellana, M. F., & Hopkins, M. (2011). Facilitating intercultural communications in parent-teacher conferences: Lessons from child translators. Multicultural Perspectives, 13, 148-154. doi: 10.1080/15210969.2011.594387

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. Psychological Assessment, 6(4), 304-312. doi: 10.1037/1040-3590.6.4.304

Gudmundsson, E. (2009). Guidelines for translating and adapting psychological instruments. Nordic Psychology, 61(2), 29-45. doi: 10.1027/1901-2276.61.2.29

Gummere, F. B. (1910). Beowulf, translated. In C. W. Eliot (Ed.). The Harvard Classics, Vol. 49. New York: P.F. Collier & Son. Retrieved from https://legacy.fordham.edu/halsall/basis/beowulf.asp

Hallberg, G. R. (1996). Assessing bilingual and LEP students: Practical issues in the use of interpreters. NASP Communique, 25(1), 16-18.

Hambleton, R. K. (2005). Issues, designs, and technical guidelines for adapting tests into multiple languages and cultures. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.). Adapting educational and psychological tests for cross-cultural assessment. (pp. 3-38). Mahway, NJ: Lawrence Erlbaum Associates.

Hui, C. H., & Triandis, H. C. (1989). Effects of culture and response format on extreme response style. Journal of Cross-Cultural Psychology, 20, 296–309. doi: 10.1177/0022022189203004

International School Psychology Association (ISPA, 2011). Code of ethics. Retrieved from http://www.ispaweb.org/wp-content/uploads/2013/01/The_ISPA_Code_of_Ethics_2011.pdf.

International Test Commission (ITC, 1999). International test commission guidelines for translating and adapting tests. Johannesburg, South Africa: Author.

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

Kankaraš, M., & Moors, G. (2010). Researching measurement equivalence in cross-cultural studies. Psihologija, 43(2), 121-136. doi: 10.2298/PSI1002121K

Klaeber, F. (Ed.) (1922). Beowulf and the fight at Finssburg. Boston: D.C. Health & Co. Retrieved from https://legacy.fordham.edu/halsall/basis/beowulf-oe.asp

Krach, S. K., Doss, K. M., & McCreery, M. P. (2015). Multilingual versions of popular social, emotional, and behavioral tests: Considerations for training school psychologists. Trainer's Forum: Journal of the Trainers of School Psychologists, 33(3), 3-26.

Kumman, A. J. (2005). Towards a model of test evaluation: Using the Test Fairness and the Test Context Frameworks. In L. Taylor & C. J. Weir (Eds.). Multilingualism and assessment: Achieving transparency, assuring quality, sustaining diversity. Proceedings of the ALTE Berlin Conference. (pp. 239-251). Cambridge, NY: Cambridge University.

Lynch, E. W., & Hanson, M. J. (1996). Ensuring cultural competence in assessment. In M. McLean, D. B. Bailey, Jr., & M. Wolery (Eds.), Assessing infants and preschoolers with special needs, 2nd edition (pp. 69-94). Englewood Cliffs, NJ: Prentice Hall.

National Association of School Psychologists. (NASP, 2010). Principles for professional ethics. Bethesda, MD: Author.

Ochoa, S. H., Gonzalez, D., Galarza, A., & Guillemard, L. (1996). The training and use of interpreters in bilingual psycho-educational assessment: An alternative in need of study. Diagnostique, 21, 19-40. doi: 10.1177/073724779602100302

Ochoa, S. H., Riccio, C., Jimenez, S., de Alba, R. G., & Sines, M. (2004). Psychological assessment of English language learners and / or bilingual students: An investigation of school psychologists' current practices. Journal of Psychoeducational Assessment, 22(3), 185-208. doi: 10.1177/073428290402200301

Oxford Dictionary (2015). Word of the Year: 2015. Oxford, UK: Oxford University Press. Retrieved from http://blog.oxforddictionaries.com/2015/11/word-of-the-year-2015-emoji/

Paone, T. R., Malott, K. M., & Maddux, C. (2010). School counselor collaboration with language interpreters: Results of a national survey. Journal of School Counseling, 8(13), 1-30.

Ren, X. S., Amick, B., Zhou, L., & Gandek, B. (1998). Translation and psychometric evaluation of a Chinese version of the SF-36 Health Survey in the United States. Journal of Clinical Epidemiology, 51(11), 1129-1138. doi: 10.1016/S0895-4356(98)00104-8

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

Roggen, V. (2014). Expanding the area of classical philology: International words. Nordlit, (33), 321-328. doi: 10.7557/13.3166

Sattler, J. (2008). Assessment of children: Cognitive foundations (5th ed.). San Diego, CA: Author.Sattler, J. M., Oades-ese, G. V., Kitzie, M., & Krach, S. K. (2014). Chapter 4:

Culturally and linguistic diverse children. In J. Sattler, & R. D. Hogue (Eds.), Assessment of Children: Behavioral and Clinical Applications (4th ed.) (pp. 125-159). San Diego, CA: Author.

Solano-Flores, G., Trumbull, E., & Nelson-Barber, S. (2002). Concurrent development of dual language assessments: An alternative to translating tests for linguistic minorities. International Journal of Testing, 2(2), 107-129. doi: 10.1207/S15327574IJT0202_2

Sotelo-Dynega, M., & Dixon, S. G. (2014). Cognitive assessment practices: A survey of school psychologists. Psychology in the Schools, 51(10), 1031-1045. doi: 10.1002/pits.21802

Swender, E. (2003). Oral proficiency testing in the real world: Answers to frequently asked questions. Foreign Language Annals, 36, 520–526. doi: 10.1111/j.1944-9720.2003.tb02141.x

Tse, L. (1995). Language brokering among Latino adolescents: Prevalence, attitudes, and school performance. Hispanic Journal of Behavioral Sciences, 17, 180–193. doi: 10.1177/07399863950172003

U. S. Office of Minority Health (2001). Eliminating racial and ethnic disparities in health. Washington, DC: U. S. Department of Health and Human Services.

Van der Velde, M. E., Feij, J. A., & van Emmerik, H. (1998). Change in work values and norms among Dutch young adults: Ageing or societal trends? International Journal of Behavioral Development, 22(1), 55-76. doi: 10.1080/016502598384513

Van Teijlingen, E., & Hundley, V. (2002). The importance of pilot studies. Nursing Standard, 16(40), 33-36. doi: 10.7748/ns2002.06.16.40.33.c3214

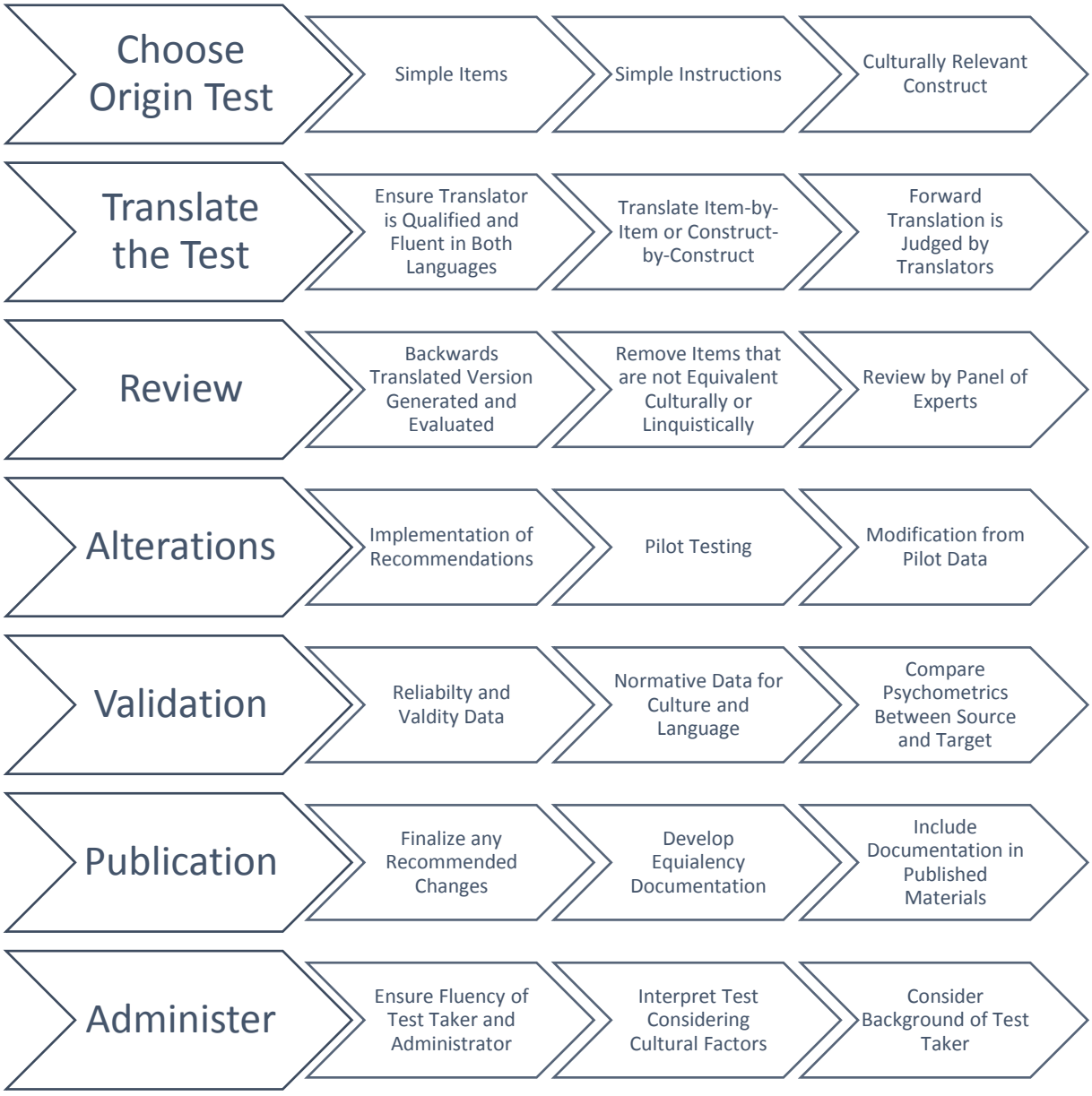
Van de Vijver, F. J. R., & Hambleton, R. K. (1996). Translating tests: Some practical guidelines. European Psychologist, 1(2), 89-99. doi: 10.1027/1016-9040.1.2.89

Van de Vijver, F. J. R., & Leung, K. (1996). Methods and data analysis of comparative research. In J. W. Berry, Y. H. Poortinga, & J. Pandey (Eds.), Handbook of cross-cultural psychology. Vol. 1: Theory and method, 2nd Edition (pp. 257-300). Boston: Allyn & Bacon.

Van de Vijver, F. J., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. European Journal of Psychological Assessment, 13(1), 29-37. doi: 10.1027/1015-5759.13.1.29

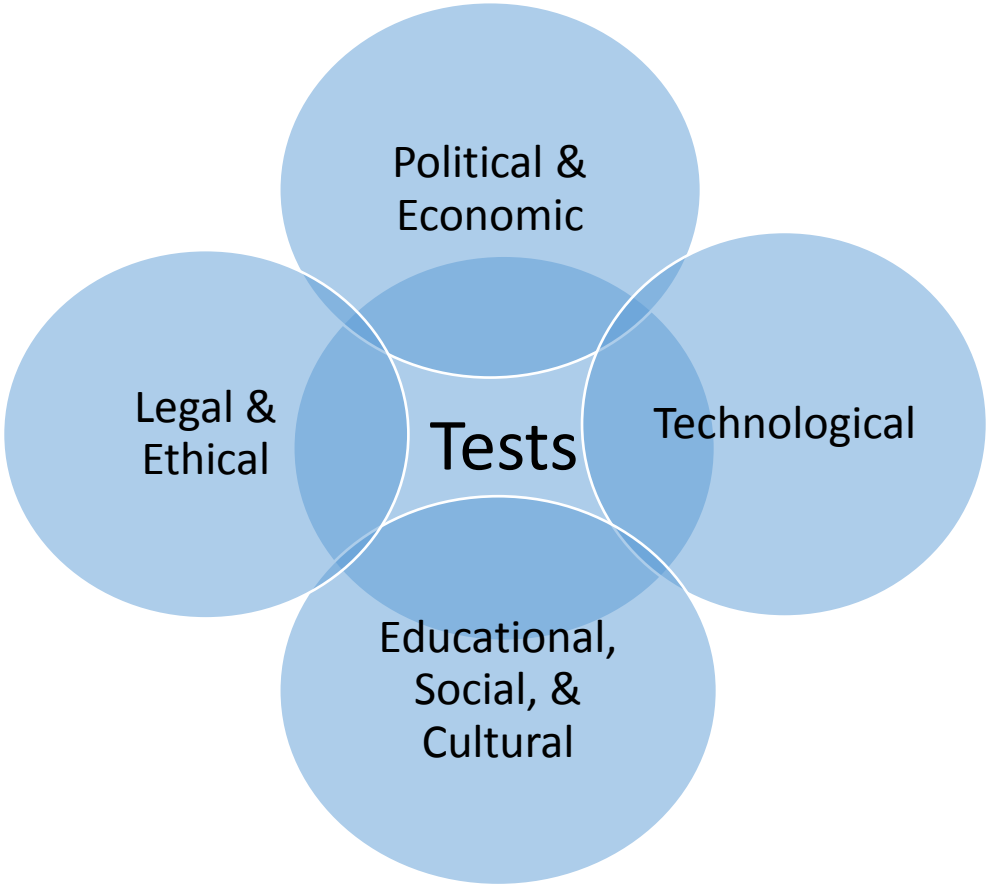
Van de Vijver, F., & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. Revue Européenne de Psychologie Appliquée / European Review of Applied Psychology, 54(2), 119-135. doi: 10.1016/j.erap.2003.12.004

Figure 1: Steps to Test Translation



Note. This table was adapted from the writings of Bracken, B. A., & Barona, A. (1991) as well as Butcher (1996) and Geisinger (1994).

Figure 2. Test Context Framework



Note. This figure was copied with permission from Kumman (2005), p. 241.

Table 1: Organizational Guidelines for Translators and Translating Tests

Authors	Date	Standard Description
AERA, APA, NCME	1999	Interpreters used in assessments should be • experts in translating • fluent in the original and target language • have a basic understanding of the assessment process
APA	2010	Psychologists • take into account culture and language in test interpretation • ensure consent for testing is without linguistic or cultural bias • are knowledgeable about cultural or linguistic differences
NASP	2011	School psychologists should • ensure that consent for testing is understandable taking into consideration the language and culture of the client • practice in a non-discriminatory manner regarding individuals who are linguistically different • conduct fair assessments taking into consideration culture and language
ISPA	2011	Use these steps when choosing someone to work with linguistically diverse clients • first, identify a school psychologist who speaks the language • next, use a knowledgeable colleague who speaks the language • finally, bring in a properly prepared translator. School psychologist are responsible to ensure that the translator • be prepared • translate with accuracy • maintain client confidentiality
ITC	2005	Ensure score equivalence across culturally and linguistically diverse groups by having test developers • verify cultural and linguistic differences when adapting a test • write test materials (e.g., handbooks, directions, etc.) to include any language issues related to the intended population • ensure that test procedures used are familiar to all populations • present evidence (including statistical evidence) that ensures and documents equivalency across all language versions • consider content validity ensuring items meets standards for cultural / linguistic equivalency • offer test instructions in the original and translated languages • document any changes from one translated version to another (including evidence of equivalence and validation) • provide information on the influence of socio-cultural and ecological context when interpreting scores

Note. American Education Research Association (AERA), American Psychological Association (APA), International Test Commission (ITC), International School Psychology Association (ISPA), National Association of School Psychologists (NASP), National Council on Measurement in Education (NCME).

Note. This table was copied with permission from the Trainers of School Psychology Forum (TSP, 2015).

Table 2: Overview of kinds of bias and their possible causes.

Kind of Bias	Source
Construct	– incomplete overlap of definitions of the construct across cultures – differential appropriateness of item content (e. g., skills do not belong to the repertoire of either cultural group) – poor sampling of all relevant behaviors (e. g., short instruments covering broad constructs) – incomplete coverage of the psychological construct
Method	– differential social desirability – differential response styles such as extremity scoring and acquiescence – differential stimulus familiarity – lack of comparability of samples (e. g., differences in educational background, age, or gender composition) – differences in physical testing conditions – differential familiarity with response procedures – tester effects – communication problems between subject and tester in either cultural group
Item	– poor item translation – inadequate item formulation (e. g., complex wording) – one or a few items may invoke additional traits or abilities – incidental differences in appropriateness of the item content (e. g., topic of item of educational test not in curriculum in one cultural group)

Note: Copied with permission from van de Vijver, F. J., & Poortinga, Y. H. (1997). Towards an integrated analysis of bias in cross-cultural assessment. European Journal of Psychological Assessment, 13(1), pg. 34.