A Synopsis of the International Test Commission's Guidelines for Translating and Adapting Tests: Part II

Eric Gehrig, PhD*
Vice President
Research & Development
Target Training International, Ltd.

Ron Bonnstetter, PhD†
Senior Vice President
Research & Development
Target Training International, Ltd.

February 13, 2020

Abstract

This is the second in a series of monthly articles outlining the International Test Commission's (ITC) Guidelines for Translating and Adapting Tests. The ITC guidelines are provided in a six-category format that includes the topics pre-condition, test development, confirmation, administration, scoring and interpretation, and documentation. It is the goal of this series to provide the motivation behind and an outline for the TTI Success Insights translation protocol.

1 Introduction

The International Test Commission's (ITC) stated goal is to assist in the exchange of information on test development and use among its members and affiliate organizations as well as with nonmember societies, organizations, and individuals who desire to improve test-related practices, see [4]. In support of this goal, the ITC has developed a series of guidelines ranging across a spectrum that includes comments on test use, quality control, test takers, and using tests for research, to name a few.

One of the major foci of the guidelines on test translation and adaptation, see [3], is to educate on the differences between translation and adaptation. The ITC states the following:

> Test translation is probably the more common term, but adaptation is the broader term and refers to moving a test from one language and culture to another. Test adaptation refers to all of the activities including: deciding whether or not a test in a second language and culture could measure the same construct in the first language; selecting translators; choosing a design for evaluating the work of test translators (e.g., forward and backward translations); ...

The guidelines are organized into six categories

1. Pre-conditions;
2. Test Development;
3. Confirmation [Empirical Analyses];
4. Administration;
5. Score Scales and Interpretation;
6. Documentation.

This article mainly covers the third item in the above list, Confirmation.

Before going on to the guidelines, we present a short motivation for the importance of having validated test adaptation protocols. Figure 1.1 shows the emotional response of a multi-lingual individual to the word *enthusiastic* presented to the person in multiple languages. This person's

*PhD, Mathematics, Arizona State University, 2007.
†Professor Emeritus, University of Nebraska-Lincoln.

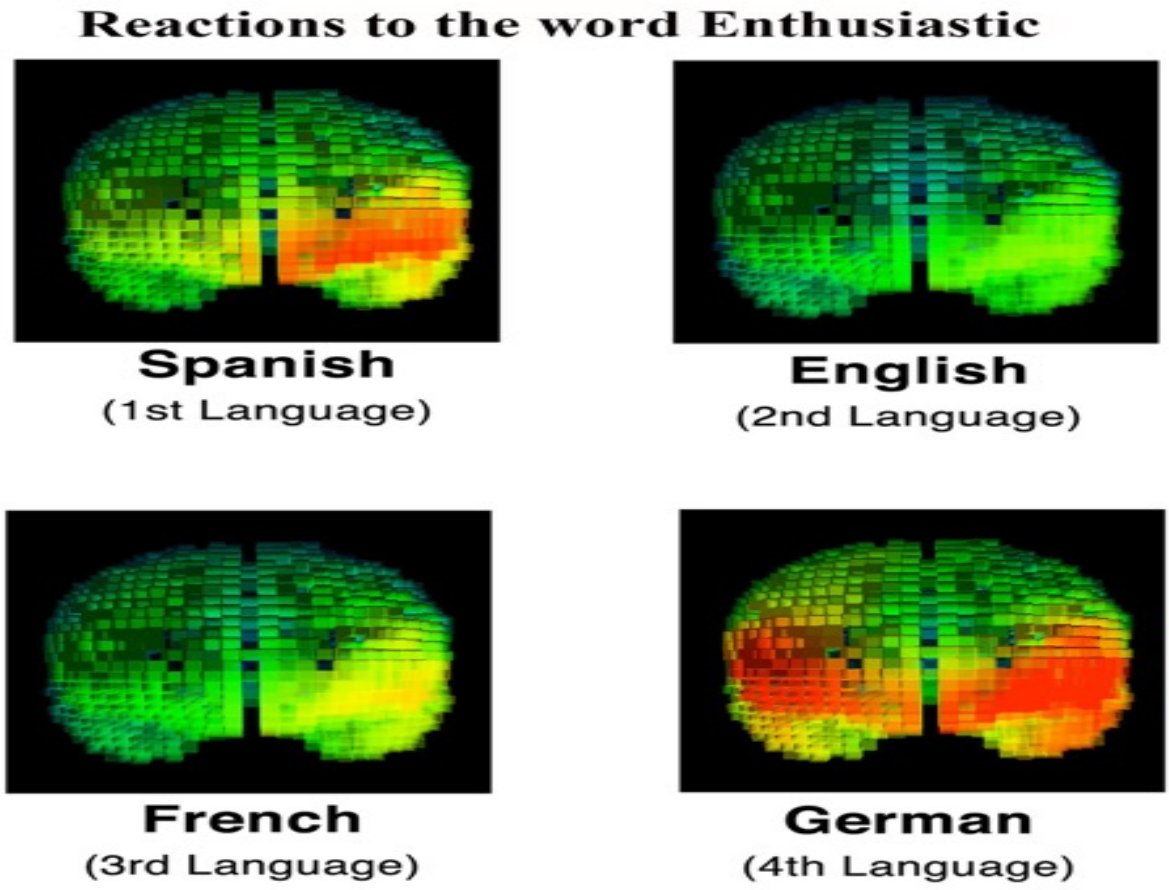

Figure 1.1: Emotional response to *enthusiastic*

native language is Spanish (Mexico). The following explanation of the brain images is taken from [2].

> She showed the maximum positive response (left-side activation) to her native language (Spanish), a relatively neutral (no overt activation in either hemisphere) to English, a moderately positive response (some left side activation) to French, and a strong and complex response (activation on both sides) to German. This was consistent with the fact that she was more comfortable with languages like her own and found German difficult and requiring effort to understand.

2 Confirmation Guidelines

While there are only four main guidelines in the Confirmation section of [3], there is an incredible amount of information and insights covered under this heading.

C-1

Select samples with characteristics that are relevant for the intended use of the test and of sufficient size and relevance for the empirical analyses.

This appears to be a bit of an obvious statement. However, the subtleties underlying the statement, and their implications, may or may not be as obvious. As noted in the first part of this series, there are some fairly strict guidelines on how to translate or adapt a test from

one language to another. It would seem that if we require the target language translator to be native to the language and the culture, then the test samples should also be native to the language and the culture. In other words, if the target language is Spanish from the Americas, it would not be acceptable to take samples from Spain or to find a Spanish speaking population in the U.S.

[3] also makes some interesting comments regarding reliability studies. We provide the following quote from pp. 17:

> However, it is often overlooked that reliability is a joint characteristic of the test and the population (McDonald, 1999) - because it depends on both the true score variance (population characteristic) and error variance (test characteristic). Therefore, the same error variance can lead to a higher reliability simply due to the larger true score variance in thee source language group. (McDonald, 1999) shows that the Standard Error of measurement (which is the square root of error variance) is in fact a more appropriate quantity to compare between samples, not reliability.

It is noted that the reference to (McDonald, 1999) is not provided in the attached bibliography. At the time of writing this article, the source [3] unfortunately does not provide the full reference in its bibliography. The authors of this article are researching to find the exact source.

The Guidelines provide the following rough guidance on minimal sample sizes:

1. Identification of potentially biased test items require a minimum of 200 persons per version of the test;
2. Item Response Theory (IRT) analyses require a minimum of 500 persons per version of the test;
3. Factor analytic studies require a minimum of 300 persons per version of the test.

The authors of this article do note that in many studies conducted by the TTI research team have shown that samples of 1000 still show some signs of instabilities in the convergence of IRT models in Likert-style response formats. As such, we typically require a minimum of 1200 persons per version of the test. It is also noted that in some cases this is a very difficult level of participation to achieve and limited information may be available for smaller sample sizes.

C-2

Provide relevant statistical evidence about the construct equivalence, method equivalence, and item equivalence for all intended populations.

Before discussing the area of item equivalence, note that the approaches for construct and method equivalence are discussed in part one of this series of articles under guidelines **PC-2** and **PC-3**, respectively.

Item equivalence is essentially studied under the name *differential item functioning (DIF) analysis*. According to [3],

> In general, DIF exists if two test-takers, from two different (cultural-linguistic) populations, have the same level of teh measured trait but have different response probability on a test item.

From the perspective of the TTI research team, one can replace the statement "different (cultural-linguistic) populations" with "categories of populations" in the sense that in the U.S. we have a need to study the existence of *disparate impact* assessment items may have on protected classes (categories) such as race, gender, age, and so on.

[3] points to two specific areas as potential sources of DIF: translation problems and cultural differences. As an example of a translation problem, it is necessary to avoid the use of vocabulary in the target language that is correct in meaning but not commonly used in a subset of the target language local culture. The following is a partial list of potential problem areas:

1. Change in item difficulty due to sentence length, complexity, or use of easy or difficult vocabulary;
2. Change of meaning due to deletion of sentences, inaccurate translation, more than one meaning in target language;
3. Non-equivalent meaning of words across cultures;
4. Cultural differences causing items to function different across languages.

[3] provides the following guidance on procedures to evaluate DIF.

1. IRT based methods;
2. Mantel-Haenszel (MH) procedure and extensions;
3. Logistic regression (LR) based procedures;
4. Restricted factor analytic (RFA) procedures.

While it was noted earlier that construct equivalence is covered, in part, in **PC-2**, [3] states the following additional information about construct equivalence:

> Finally, yet importantly, this guideline will require researchers to address construct equivalence. There are at least four statistical approaches for assessing construct equivalence across source and target language versions of a test: Exploratory factor analysis (EFA), confirmatory factor analysis (CFA), multidimensional scaling (MDS), and comparison of nomological networks (Sireci, Patsula, & Hambleton, 2005).

The reference for (Sireci, Patsula, & Hambleton, 2005) is provided in [5]. It is noted in [3] that there are no commonly agreed-upon rules for deciding when structures can be considered equivalent. Hence, as EFA does not accommodate separate factor structures, it is more desirable to use CFA, weighted MDS (WMDS), or Exploratory Structural Equation Modeling (ESEM).

C-3

Provide evidence supporting the norms, reliability, and validity of the adapted test version of the test in the intended populations.

Directly from [3]:

> The norms, validity evidence, and reliability evidence of a test in its source language version do not automatically apply to other possible adaptations of the test in to different cultures and languages.

Further, [3] suggests essentially following the standards set forth in [1] as well as in [6]. These standards state that one should provide evidence of validity based on the following five areas:

1. Test content;
2. Internal structure;
3. Relations to other variables;
4. Response processing;
5. Consequences of testing.

[3] also provides a short list of approaches to internal structure validity evidence such as EFA, CFA, structural equation modeling (SEM), and multitrait-multimethod analyses.

C-4

Use an appropriate equating design and data analysis procedures when linking score scales from different language versions of a test.

This topic may be more readily understood in the language of parallel form tests. For example, given a high stakes testing environment, there may be an increased risk of cheating. In such cases, multiple versions of the test may be produced to lessen the probability of an individual successfully cheating. However, how does one know if a given score on, say, Form A is equivalent to that same score on Form B? It may be the

case that one form or the other is more or less difficult than its counterpart(s). As an example, it may be the case that a score of 70% on Form A is equivalent to a score of 60% on Form B due to an increased level of difficulty on Form B.

The process of determining the equivalence of scores in this example is called equating scores. If we are looking at different adaptations of a test into other languages and/or cultures, the process is called linking the scores. In order to establish the linking process, [3] suggest is it best to address the following three questions:

1. Is there evidence that the same construct is being measured in the source and target language versions of the test?/Does the construct have the same relationship with other external variables in the new culture?

2. Is there strong evidence that sources of method bias have been eliminated?

3. Is the test free of potentially biased test items?

3 Summary

This article presents an outline of the Confirmation portions of the International Test Commission's Guidelines on Test Translation and Adaptation. We present these guidelines along with a short commentary on how we are set to interpret these guidelines in preparation for a full test adaptation protocol implementation. A forthcoming article addresses the next series of guidelines on Administration.

References

[1] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.). Standards for educational and psychological testing, 2014.

[2] Thomas F. Collura, Nancy L. Wigton, Carlos Zalaquett, SeriaShia Chatters-Smith, and Ronald J. Bonnstetter. The Value of EEG-Based Electromagnetic Tomographic Analysis in Human Performance and Mental Health. *Biofeedback*, 44(2):58–65, 2016.

[3] International Test Commision. The ITC Guidelines for Translating and Adapting Tests (Second Edition). https://www.InTestCom.org, 2010.

[4] Thomas Oakland, Ype H. Poortinga, Justin Schlegal, and Ronald K. Hambleton. International test commission: Its history, current status, and future directions. *International Journal of Testing*, 1(1):3–32, 2001.

[5] S.G. Sireci, L Patsula, and R.K. Hambleton. Statistical Methods for Identifying Flaws in the Test Adaptation Process. In R.K. Hambleton, P. Merenda, and C. Spielberger, editors, *Adapting Educational and Psychological Tests for Cross-cultural Assessment*, pages 93–116. Lawrence Erlbaum Publishers, Mahwah, NJ, 2005.

[6] Stephan G. Sireci and Tia Sukin. Test Validity. In Kurt F. Geisinger, Bruce A. Bracken, Janet F. Carlson, Jo-Ida C. Hansen, Nathan R. Kuncel, Steven P. Reise, and Michael C. Rodrigues, editors, *APA Handbook of Testing and Assessment in Psychology Volume 1: Test Theory and Testing and Assessment in Industrial and Organizational Psychology*, chapter 4, pages 61–84. American Psychological Association, 750 First Street NE, Washington, D.C. 20002-4242, 2013.