SIOP White Paper Series

SOCIETY for INDUSTRIAL and ORGANIZATIONAL PSYCHOLOGY

SCIENCE FOR A SMARTER WORKPLACE

Test Adaptation and Measurement Equivalence

Neal Schmitt and Dragos Iliescu

A White Paper prepared by the Visibility Committee of the Society for Industrial and Organizational Psychology.
440 E Poe Rd, Suite 101 Bowling Green, OH 43402

Correspondence regarding this whitepaper should be addressed to either Neal Schmitt (schmitt@msu.edu) or Dragos Iliescu (dragos.iliescu@fpse.unibuc.ro)

Table of Contents

Authors

Neal Schmitt
Michigan State University

Neal Schmitt is Emeritus Professor of Psychology and Management at Michigan State University. He was editor of *Journal of Applied Psychology* from 1988-1994 and has served on a dozen editorial boards. He has received the Society for Industrial and Organizational Psychology's Distinguished Scientific Contributions Award (1999) and its Distinguished Service Contributions Award (1998), as well as several other career awards from the American Psychological Association and the Academy of Management. In 2014, he was named a James McKeen Cattell Fellow of the American Psychological Society. He served as the Society's president in 1989-90 and as the president of Division 5 of APA (Measurement, Evaluation, and Statistics). He has authored three textbooks, coedited four books, and has published approximately 250 peer-reviewed papers and chapters. His current research centers on the effectiveness of organizations' selection procedures, college admissions processes, and the outcomes of these procedures.

Dragoș Iliescu
University of Bucharest

Dragoș Iliescu is a professor of Psychology with the University of Bucharest. He has been active as a consultant for the past 20 years, being involved in and having led important projects related to tests, testing, and assessment (among them more than 100 test adaptation projects), mainly in South-Eastern Europe but also in South-East Asia, Africa, the Middle East and South America. Dragoș Iliescu has served in various capacities for a number of national and international professional associations; he is the current president (2016-2018) of the International Test Commission (ITC). He is an associate editor for the *European Journal of Psychological Assessment,* and the author of over 100 scientific papers, book chapters, and books, among them the coeditor of the acclaimed *ITC International Handbook of Testing and Assessment,* published in 2016 by Oxford University Press, and the author of an important monography *(Adapting Tests in Linguistic and Cultural Situations)* published with Cambridge University Press.

Introduction

Test adaptation is a comprehensive scientific process by which a measure is transformed for appropriate usage from an original language and culture to a target language and culture. Test adaptation is based on translation but may also encompass transformations to other components of the test, such as item content, item format, scaling of items, structure of the test, scoring keys, norms, graphical layout, and others. Test adaptation is a ubiquitous process and is present either explicitly or implicitly in most of psychological research and practice. .

Although published overwhelmingly in English language outlets, much if not most research on adaptation is conducted in countries where the native language is not English, and data are collected with adapted forms of the reported measures. Although tests are preferentially developed in some few countries and usually in the English language, they are used all around the world by professionals in their work with clients. Even within English-speaking countries, tests are frequently adapted to reflect differences in user background and experiences. The basic assumption behind using the adapted measure is that it is, for all practical reasons, similar to the original. A violation of this assumption has important implications (van de Vijver & Poortinga, 2005). Research published with various different-language forms of a measure is only generalizable if these same-named but different-language or cultural measures are similar to each other. If the adapted form of a test is not similar to the original, research conclusions derived from the new version do not generalize, and validity evidence based on the original test may not be applicable. Practical interpretations and decisions made on the basis of an adapted test are only warranted if the validity evidence accumulated on the original test applies or transfers to the adapted form.

The assumption that the original and adapted measure are similar reflects a form of validity generalization, that is, the degree to which evidence of validity obtained in one situation or context can be generalized to other situations or contexts without the need to explicitly research the validity in the new situation or context (AERA, APA, & NCME, 2014). A test is only adapted for professional use because of its usefulness, which it is warranted by the validity evidence that the test has amassed in time and which is contained in its interpretive documentation (e.g., technical manuals, user guides, white papers, intervention planners, etc.). If the evidence and interpretive documentation of a test cannot be "inherited" and used in the new culture, then developing a new test may be more attractive than adapting one.

The similarity between the original form of a test and its adapted version is called "equivalence" or "invariance." When a test lacks equivalence, it is usually considered biased when used in contexts or with groups for which the test has exhibited nonequivalence.

Equivalence should be understood as one aspect of validity (Iliescu, 2017). The relationship between test scores (and the underlying constructs the respective test scores should reflect) is the same across different uses of the test. These different instances can be, for example, different groups (i.e., equivalence of test scores for minority and majority

groups), contexts (i.e., equivalence of test scores for low-stake vs. high-stake testing), or separate forms of a test (i.e., original source-language form vs. adapted target-language form). This latter case of equivalence is important for test adaptations.

The most widely accepted definition of validity is that "validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (AERA, APA, & NCME, 2014, p.11). In practical settings, the interpretation of test scores provided by an adapted form of a test is intended to follow the evidence provided by the original test. In research settings, the interpretation of test scores stemming from an adapted form of a test is intended to contribute to the evidence accumulated by the original test. Such an interpretation of scores should not be given unless the two forms are virtually identical. Equivalence therefore refers to the degree to which the empirical evidence supports the fact that the adapted version of the test is similar to the original version and warrants the same score interpretations.

It should be noted that equivalence, as traditionally discussed in the literature, is concerned foremost with measurement aspects, that is, with the similarity of the various components of the test (e.g., items, scales, structure), and does not actually offer evidence for similarity of relationships between test scores and external criteria in the two cultures or groups. Such differential prediction is usually indicated by slope and intercept differences across subgroups when a criterion is regressed on test scores (Bartlett, Bobko, Mosier, & Hannon, 1978).

When compared against the six sources of validity evidence outlined by the Standards for Educational and Psychological Assessment (AERA, APA, & NCME, 2014, pp. 13-16) equivalence may offer positive evidence on only three: (a) evidence based on test content (i.e., is the content of the original and adapted measure equivalent?), (b) evidence based on response processes (i.e., the cognitive processes in which test takers engage are similar), and (c) evidence based on internal structure (i.e., the relationships between the different components of the test, such as test items and test scales, are similar). As a result, equivalence does not offer strong evidence for generalizability: It only offers some evidence for the possibility to generalize from one cultural context to another, exclusively related to measurement aspects. Hence, also the preferred wording of "measurement equivalence." A great deal of the literature on equivalence involves investigations of the internal structure of the test, usually through item response theory or confirmatory factor analytic examinations of data collected from multiple groups of respondents. A measure is invariant when members of different populations who have the same standing on the construct being measured receive the same score on the test. Conversely, a measure lacks equivalence when two individuals from different populations who are identical on the construct score differently on the test.

A measure is invariant when members of different populations who have the same standing on the construct being measured receive the same score on the test.

Item response theory. Two major methods have been used to identify invariance. The first arising from IRT posits that test items are equivalent when the curve representing the relationship between the underlying trait and the probability of a correct response is identical across subgroups of individuals (Embretson & Reise, 2000). Uniform DIF occurs when the expected scores of one group are uniformly higher than those of another group. Nonuniform DIF occurs when scores are lower (or higher) than expected as a function of their level on the construct being measured. Tests of significance of DIF are available as are effect sizes, and there is also a cumulative index of DIF called differential test functioning (DTF).

Confirmatory factor analysis. The second approach employs confirmatory factor analysis (CFA) to assess measurement equivalence. In factor analytic terms, each item or indicator of a latent construct comprises variance related to the underlying factor (the factor loading represents this linear relationship), unique variance unrelated to the factor, and a constant or intercept. The most basic level of invariance, termed configural invariance, is represented by a situation in which items load on the same factor across multiple groups, but the degree of relationship, represented by the factor loadings, may vary across groups. This baseline model is usually compared with a metric invariance model in which the factor loadings for each group are constrained to equality. Factor loadings represent the strength of the relationships between items and factors or, in a regression sense, the weights obtained by regressing the items on the factor they are thought to represent. When factor loadings are equal, the unit of measurement is equal across groups and cross-group predictive relationships (relationships with variables external to the factor model) are comparable. Scalar invariance requires that the intercepts, as well as the factor loadings, associated with item–factor relationships be equal across groups. With scalar invariance, the observed means of different groups can be compared meaningfully. The fourth form of measurement invariance is the invariance of the uniquenesses associated with each item. Invariance of uniquenesses indicates equivalence in the precision of measurement of each item. Some maintain that strict factorial invariance including metric, scalar, and uniqueness invariance is required for valid comparisons of observed group means. Equivalence in uniquenesses is not required for comparisons of latent means as the measurement errors are partialed out in CFA analyses (Meredith, 1993). Tests of the equality of factor variances and covariances provide evidence of the equivalence of construct relationships. Test of the equality of latent means—which would be conducted next—may often be the central research question.

As mentioned above, when some items are invariant whereas others lack equivalence, we have a condition referred to as partial invariance. Partial invariance can be incorporated into CFA models. Statistical tests of invariance follow a sequence of steps outlined by Vandenberg and Lance (2000) as well as others.

Differences across groups occur relatively frequently but are not large, often do not replicate, and are often not easily explained post hoc.

Both IRT and CFA examinations of measurement equivalence provide statistical tests of the differences across groups as well as goodness-of-fit measures for various models and may provide evidence of the degree to which a test has been successfully adapted if scores on the original and adapted version of the test are available for some group of respondents. CFA can also be used to test the various levels of equivalence as outlined above. Tests of uniform and nonuniform DIF in IRT analyses are analogous to tests of factor and scalar equivalence in CFA

Lack of equivalence can be the result of bias in a number of components of the test. These various sources of bias have been traditionally discussed in the literature under three large headings: construct bias, method bias, and item bias (van de Vijver, 2016). These sources of bias define corresponding categories of equivalence: construct equivalence, method equivalence, and item equivalence. Construct bias refers to bias related to the measured construct itself. Nonequivalence (or lack of configural invariance in CFA terms) may appear in construct-related issues if the construct that was initially targeted by the original version of the test does not exist at all or does not exist in the same way in the target culture of the adaptation process. Method bias refers to bias related to the method (the testing process) and is in fact a generic term for any number of nuisance factors that are related to the direct testing process and to the sample (sample bias), the measure itself (instrument bias), and the administration procedure (administration bias). Item bias refers to bias related to one or more items of the test. Item bias may appear due to, for example, poor translation or poor cultural adaptation. Item equivalence is the situation in which each item of the adapted form of the test elicits the same response and at the same intensity when administered in the target culture (scalar equivalence in CFA terms), as that particular item does in the original form of the test when administered in the source culture.

Differences across groups occur relatively frequently but are not large, often do not replicate, and are often not easily explained post hoc. Our recommendation is that researchers should continue to analyze responses to their measures for a lack of invariance whenever possible and consider available effect sizes and the sensitivity of the tests for invariance (Meade, 2010; Meade, Johnson, & Braddy, 2008). They should also focus on the degree to which findings of a lack of invariance are interpretable and have relevance for decisions that are made using the instrument. In addition, the impact of a lack of invariance on the evaluation of substantive hypotheses regarding relationships among variables should be examined.

References

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.

Bartlett, C. J., Bobko, P., Mosier, S. B., & Hannon, R. Jr. (1978). Testing fairness with a moderated multiple regression strategy: An alternative to differential analysis. *Personnel Psychology, 31,* 233-242.

Embretson, S. E., & Reise, S. P., (2009). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Iliescu, D. (2017). *Adapting tests in linguistic and cultural situations.* Cambridge, UK: Cambridge University Press.

Meade, A. W. (2010). A taxonomy of effect size measures for the differential functioning of items and scales. *Journal of Applied Psychology, 95,* 728-743.

Meade, A. W., Johnson, E. C., & Braddy, P. W. (2008). Power and sensitivity of alternative fit indices in tests of measurement invariance. *Journal of Applied Psychology, 93,* 569-592.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 53,* 525-543.

van de Vijver, F. J. R. (2016). Test adaptations. In F. T. L. Leong, D. Bartram, F. M. Cheung, K. F. Geisinger, & D. Iliescu (2016). *The ITC international handbook of testing and assessment* (pp. 364-376). Oxford, UK: Oxford University Press.

van de Vijver, F. J. R., & Poortinga, Y. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting educational and psychological tests for cross-cultural assessment* (pp. 39-64). Mahwah, NJ: Erlbaum.

Vandenberg, R. J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3,* 4-70. doi:10.1177/109442810031002