A Theoretical Suggestion on Testing Measurement Invariance in Adapting Parametric Measurement Tools

Gökhan İskifoğlu
Department of Classroom Teaching
European University of Lefke, TRNC
giskifoglu@eul.edu.tr

Abstract
This research paper investigated the importance of conducting measurement invariance analysis in developing measurement tools for assessing differences between and among study variables. Most of the studies, which tended to develop an inventory to assess the existence of an attitude, behavior, belief, IQ, or an intuition in a person's characterological profile, ignored testing measurement invariance for equivalency between comparable variables. With this finding, measurers lack in true validity and reliability and suffer methodological bias and have little or no chance to figure out the true differences between variables being studied. This article, therefore, explains the necessity and use of measurement invariance analysis when a researcher wanted to develop a new measurement tool or adapt a tool from one source language to another target language. The types of measurement invariance levels mentioned in this study are configural-invariance, scalar invariance, metric invariance and structural invariance analysis. The approaches used to conduct those invariance models and the way they have been interpreted were all discussed in great detail with a robust collection of supportive literature.

Introduction

The variations between groups formed/formed according to criteria like culture, gender, class level, and socioeconomic status are explored in terms of unique psychological structures in various research studies undertaken in the domains of Educational Sciences and Psychology (Lucke, 2005; Gödelek, 2005). The purpose of these investigations is to determine whether there are any variations in the psychological structures of interest between the groups listed, and if so, to draw conclusions regarding the kind and extent of those differences. Group comparisons in these studies primarily rely on measurements (observed scores) associated with the pertinent psychological structure (Wicherts, 2007). Nevertheless, proof that the pertinent metrics possess appropriate psychometric qualities is necessary for the validity of these comparisons to hold. This data frequently bolsters the assumption that a measurement, like Yi, represents an underlying latent structure, like η, within the framework of the Classical Test Theory (CTT) (Vandenberg & Lance, 1998;Vandenberg & Lance, 2000). Consequently, this calls for trust in a metric such as Yi as an indicator of η. In fact, this statement highlights the fact that measuring invariance analysis is a necessary step before group comparisons can be made. Measurement invariance is the formal assessment of a psychological measurement instrument's homogeneity in psychometric qualities among several groups, according to Herdman (1998). To demonstrate measurement invariance, Vandenberg and Lance (2000) provide a five-stage logical process and methodologies for hypothesis testing (Schraw, Dunkle, & Bendixen, 1995). A hypothesis concerning the degree of measurement invariance is successively examined at each of these phases (Schommer, 1990). Each model in this method is built at a certain stage using the model from the stage before it. As a result, measurement invariance at a given stage is investigated by contrasting the model's fit with the data at that stage with the model's fit from the stage before. The phases that the researchers have suggested also point to several kinds of measurement invariance (Salzberger, Sinkovics, & Schlgelmich, 1999)

1. Configural Invariance: This step involves testing a hypothesis about the equality or invariance of a psychological assessment instrument's factor structure between groups. This is known as configuration invariance. If configural invariance is proven, then the assessment instrument's items must reflect the same psychological structure across all groups (Vandenberg and Lance, 1998).(see Fifure 1A and 1B).
2. Metric Invariance: Here, a hypothesis about the equality or invariance of the factor loadings (lambda: λ), or regression slopes, of the items that make up a psychological assessment tool between groups is evaluated (Ravindran, Greene, & DeBacker, 2005). While a substantial difference implies item bias, the absence of a statistically significant difference in the factor loadings for the items between comparison groups suggests that the meanings of the items may be similar or equivalent for these groups (see Figure 1 C and 1D).
3. This step involves testing a hypothesis about the equality or invariance of the constant term (tau: τ) in the regression equations developed for the items that make up the psychological measuring tool (Mark, & Wan, 2005). This is known as scalar invariance. Equal intercepts and metric invariance are prerequisites for scalar invariance in the measuring process (see Figure 1 E and 1 F) (Vandenberg and Lance, 2000). It is said that two scores from Group A assessed in Group B must equal one another for the measurement sources to be equal. It is suggested that there might be bias if the two scores in Group A are equal to three points in Group B (Fleck, Poirier-Littre, M.F., Guelfi, Bourdel, & Loo, 1995). This bias would show up as a difference in the intercepts (Salzberger et al., 1999;

Wicherts, 2007). Two categories of item bias are defined in this regard: (1) Uniform Bias: Items in a measurement instrument that have factor loadings that are invariant between groups but different intercepts for each group are said to display uniform bias and (2) Nonuniform Bias: When a measurement tool has nonuniform bias, different groups have different factor loadings and intercepts of the item types (Eroğlu, & Güven, 2006). The low predicted amount of scores in this case, observed in the group when the τ value is relatively low, depends on both the true value of η and the τ value (Crocker & Algina, 1986; Chan, 2003).

4. Invariant Uniqueness: In this stage, a hypothesis is tested on the equality/invariance of specific variances, i.e., error terms, of the items forming the measuring instrument across comparison groups. Furthermore, it should be mentioned that this test is also regarded as a test of the invariance of the indicators' reliability if proof of the invariance of factor variances is found (Vandenberg and Lance, 2000).

5. In this phase, a hypothesis about the equality or invariance of factor variances among comparison groups is examined (Bryne, & Watkins, 2003). The purpose of this hypothesis test is to ascertain whether the conceptual structure of the construct that the psychological assessment tool is meant to measure, with equal ranks, is the same for the comparison groups with regard to how they react to its indicators (Brown, 2006; Mark and Wan, 2005).

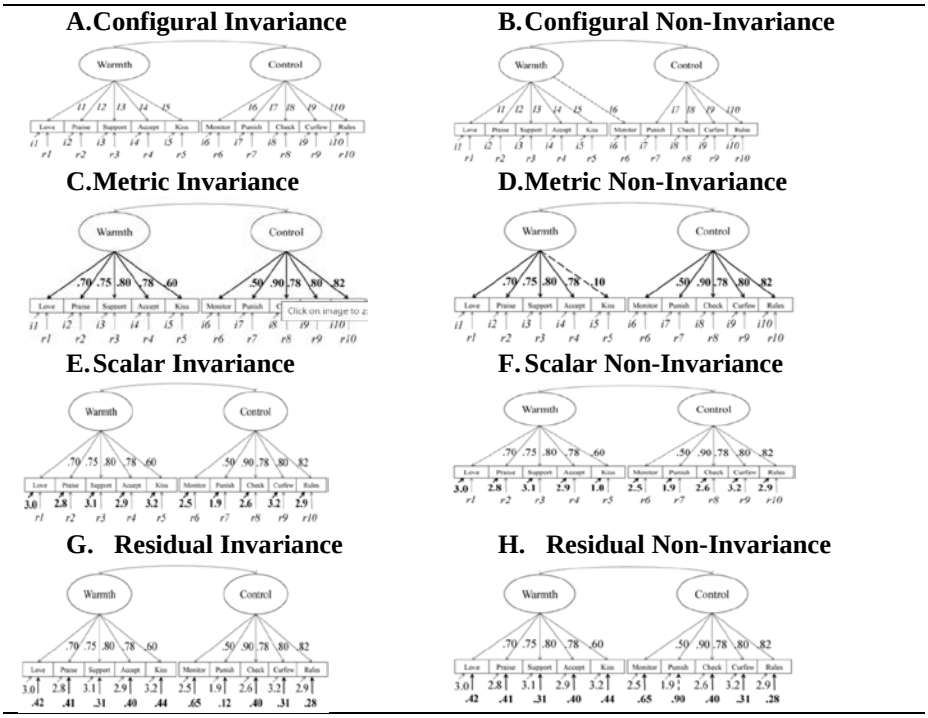

Figure 1. Invariance models in terms of their free estimation levels

Specifically, in this article, those 5 important invariance models are discussed and made invaluable contributions in to how can a study of measurement tool development study can benefit out of its suggestions. This was so crucial since the research into the related literature showed that the majority of studies conducted to develop measurement tools lack in terms of its misuse of statistical techniques and methodology (Bell, 2007). Nevertheless, most of the measurement tools developed in Turkey for assessing a psychological construct have not considered conducting measurement invariance analysis to see if the differences measured are due to true differences or differences that occurred because they were not equivalent. For this purpose, the first issue that need to be discussed in this regard will be the issue of equivalency.

Equivalency in Test Adaptation

Studies pertaining to scale adaptation often involve two primary phases: the analysis of psycholinguistic characteristics (language adaptation) and the assessment of psychometric attributes (reliability and validity). According to many sources (Aksayan and Gözüm 2002; Esin 2014; De Lima Barroso et al. 2018; International Test Commission 2018; Thammaiah et al. 2016; World Health Organization 2018), each of these stages also requires a number of procedures to be completed. These steps are covered in detail in the section that follows, and as a valuable resource, the International Test Commission's (ITC) Guide for Translating and Adapting Tests. Due to variations in conceptualization and expression, a scale's initial structure may alter while being translated into a different language. The scale items must be thoroughly examined, the required translations must be made to assure

meaning in the target language, and the scale must be standardized in accordance with the norms of those who use the target language in order to minimize this shift. The scale's psychometric (reliability and validity) scores could be poor if this process doesn't receive the necessary attention. Thus, during the translation process, much consideration should be paid to the choice of translators and the translation methodology (Aksayan and Gözüm 2002). In the current international literature, it is frequently recommended to follow the sequence of the four processes below for adapting a scale to a different language and culture and ensuring linguistic validity (Beaton et al. 2000; De Lima Barroso et al. 2018; International Test Commission 2018; Thammaiah et al. 2016; WHO 2018). Scale adaptation is the process of modifying scales created in any language and culture for use in a different language and culture (Aksayan and Gözüm 2002; Esin 2014). In addition to saving researchers time when creating a new scale, translating and utilizing internationally recognized scales into Turkish facilitates communication and yields information that is comparable to results from other societies. Furthermore, obtaining accurate, valid, and thorough measurement results can be achieved affordably and effectively by employing these standardized measuring instruments during the data collecting process (De Lima Barroso et al. 2018; Esin 2014). Additionally, obtaining accurate, valid, and thorough measurement results can be achieved affordably and effectively by employing standardized measurement instruments during the data collecting process (De Lima Barroso et al. 2018; Esin 2014). Furthermore, translating scales into Turkish could facilitate the process for researchers who lack the skills and expertise to create a scale and increase access to a wider variety of measuring tools for study. A methodical research procedure, scale adaptation can cause confusion if the researcher is unfamiliar with the topic. Consequently, training and advice on the issue may be advantageous for the researcher.

Process of Test Adaptation

Group translation

Using the group translation method, two or more people who are fluent in both languages translate the scale from the originating language into the target language. With this approach, members of the group should translate more freely in order to relieve pressure to persuade one another and come to a compromise. To prevent the use of particular jargon in the relevant sector, it is advised to consider individuals with education from a variety of fields when choosing translators. To prevent the use of particular jargon in the relevant sector, it is advised to consider individuals with education from a variety of fields when choosing translators. Members of the translation team should also have understanding of and experience with research methods, the translation process, and the local culture. Furthermore, researchers must perform follow-up interviews and take part in the writing of the final article rather than serving as translators. Then, researchers and members of the translation group come to a consensus together to create a single text (Esin 2014; International Test Commission 2018; World Health Organization 2018).

Back Translation Method

The second important phase in the language validation process is back translation, which is a suggested way to confirm that an original language translation to the target language is accurate. This phase ensures that conceptual flaws and inconsistencies are understood and acts as a quality control activity. It facilitates comprehending the semantic correspondence between the source and translated texts. The best translators for the back translation process are those who are not affiliated with the research team and are not conversant with the research's subject matter (Beaton et al. 2000; International Test Commission, 2018; World Health Organization, 2018).

Experts View on the Phenomenon Being Studied

This group consists of researchers, translators, and participants from the field who were involved in earlier investigations. The scale owner may be asked to assist in elucidating any differences (if any) that are seen between the original and target versions, or they may be asked to join the expert panel if they are fluent in the language. The adaption of scales discusses the idea of equivalency in language and meaning. The technique of conceptual equivalency is advised for adapting the scale to Turkish, as there may be difficulties in precisely translating some terms and concepts from the original scale to match the linguistic and cultural characteristics of the target language. For this reason, getting professional advice is also recommended. Conveying the same meaning with words and sentences that are culturally specific is taken into consideration in conceptual equivalency. Some elements from the original scale can be translated into the target language utilizing multiple items when it is thought essential. In this instance, the most relevant item may be selected in the item selection stage of the scale's reliability and validity assessment (Aksayan and Gözüm 2002; Esin 2014; WHO 2018).

Initial Application of the Tool Being Adapted

At this point, the scale is usually delivered to a sample of 30 to 40 members of the target audience, and the participants' thoughts and input are used to evaluate the items' acceptability and clarity. Researchers can determine whether the scale is straightforward, intelligible, appropriate for the setting, and simple using this procedure.

Furthermore, this step aids researchers in making sure that the translation is done using appropriate language and expressions that are culturally neutral (Beaton et al. 2000; International Test Commission 2018; World Health Organization 2018).

Assessment of the Psychometric Properties of the Tool Being Adapted

In the second phase of scale adaptation research, the psychometric characteristics of the scale modified for the target language—that is, its validity and reliability—must be investigated. Every measurement instrument is designed to measure a certain property precisely, under a given set of circumstances, and with respect to a particular group of people. It is inappropriate to use a measuring tool that is unable to produce precise measures or, if it does, is not acceptable for the purpose for which it was designed. As a result, it's important to take both measuring tool validity and dependability into account. The literature provides some explanations on the significance of sample size determination in reliability and validity investigations. First, the total number of items can be used to establish the sample size in validity and reliability investigations. Working with a sample size that is five to ten times the total number of items is often advised when using this strategy (Esin 2014). Conversely, validity refers to how well a measurement instrument fulfills its intended use and specifies what and how precisely/accurately it measures (Erkuş 2003; Esin 2014). A scale's validity and reliability can be ascertained using a variety of techniques. Table 2 lists these techniques. It is generally advised to utilize at least two methods for each purpose when verifying the validity and reliability of a measurement tool (Erkuş 2003).

Validity and Reliability Issues

In order to determine if the scale as a whole and its sub-dimensions measure the intended domain and whether there are different concepts beyond the intended domain, scope/content validity analysis is carried out (Gözüm and Aksayan 2002). Although content validity is really a procedure that needs to be done when developing a new scale, some literature reports that it may also be used when adapting an existing scale (Gözüm and Aksayan 2002; Esin 2014; Ljungberg et al. 2014). Nonetheless, content validity in this particular setting is not expressly addressed in current international standards (De Lima Barroso et al. 2018; International Test Commission 2018; Thammaiah et al. 2016; WHO 2018). Is content validity still required if there are no items being added or withdrawn from the scale at this time? Five scale adaptation studies in nursing and five in other professions that were completed in the last few years were reviewed in order to investigate this subject.

Conclusion

When a test was adapted from one language to another, measurement invariance must be tested following the adaptation process. This is due to one important reason and it is that the true difference can only be seek between equivalent groups. If the groups are not equivalent with one another that the differences are biased differences and they do not show the truth in advance. The tests that are conducted to test the differences between variables are also limited to the degree of invariance achieved at different levels. For instance, if the measurement model did not display a metric invariance so group-wise mean comparisons cannot be conducted between groups since their distributions are not equivalent. The same is valid through all invariance models. In addition, if there is no factorial variance invariance achieved then no regression analysis should be conducted since there are non-invariant groups. Although correct procedures have been followed through adaptation process, cultural bias may occur due to many tacit knowledge that cannot be directly assessed. Measurement invariance in addition to confirmatory factor analysis need to be thought. Otherwise, they are not considered as comparable measurement models.

References

Bell, P.D. (2007). Predictors of college student achievement in undergraduate asynchronous web-based courses. Education. 127 (4), 523-533. 20 Şubat 2008'de http://web.ebscohost.com/host/pdf?vid=8hid=3&sid=79a809cb-8dbe-48d4-9e0664743b6db9bd%40sessionmgr7

Brown, T. A. (2006). Confirmatory factor analysis for applied research. New York: The Guilford Press.

Bryne, B. M. & Watkins, D. (2003). The ıssue of measurement ınvariance revisited. Journal of Cross-Cultural Psychology. 34, (2), 155-175. 10 Aralık 2004'de http://jcc.sagepub.com/cgi/reprint/34/2/155.pdf

Chan, K. (2003). Hong Kong teacher education students' epistemological beliefs and approaches to learning. Research in Education. 69, 36-50. 26 Mayıs 2007'de http://web.ebscohost.com/ehost/pdf?vid=23&hid=105&sid=4084c36f-9029-439fa5c89de957c13c3a%40sessionmgr109

Crocker, L. & Algina, J. (1986). Introduction to Classical and Modern Test Theory. Orlando: Harcourt Brace Jovanovich Inc.

Eroğlu, S. E. & Güven, K. (2006). Üniversite öğrencilerinin epistemolojik inançlarının bazı değişkenler açısından incelenmesi. Selçuk Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 16, 295-312.

Fleck, M.P., Poirier-Littre, M.F., Guelfi, J.D., Bourdel, M.C. & Loo, H. (1995). Factorial structure of the 17-item Hamilton Depression Rating Scale. Acta Psychiatr, 92, 168172. Web: http://www3.interscience. wiley.com/cgibin/fulltext/119231102/PDFSTART?CRETRY=1& SRETRY=0

Mark, B. A. & Wan, T.T.H (2005). Testing measurement equivalence in a patient satisfaction instrument. Western Journal of Nursing Research, 27 (6), 772-787. 14 Ekim 2005'de http://wjn.sagepub.com/cgi/reprint/27/6/772.pdf

Marzooghi, R., Fouladchang, M. & Shemshiri, B. (2008). Gender and grade level differences in epistemological beliefs of ıranian undergraduate students. Journal of Applied Sciences, 8 (24), 4698-4701. 16 Mart 2009'da http://scialert.net/qredirect.php?doi=jas.2008.4698.4701& linkid=pdf

Ravindran, B., Greene, B.A. & DeBacker, T.,K. (2005). Predicting preservice teachers' cognitive engagement with goals and epistemological beliefs. The Journal of Educational Research, 98 (4). 18 Aralık 2008'de http://web.ebscohost.com/ehost/pdf?vid=11&hid =7&sid= a63c1a96-d15a-4695-aa3562626af53d0f%40sessionmgr109

Salzberger, T., Sinkovics, R. R. & Schlgelmich, B. B. (1999). Data equivalence in crosscultural research: A comparison of classical test theory and latent trait theory based approaches. Australasian Marketing Journal, 7 (2), 23-38. 18 Haziran 2007'de http://www.personal.mbs.ac.uk/rsinkovics /pubs/1999-AMJ-Equiv.pdf

Schommer, M. (1990). Effects of beliefs about the nature of knowledge in comprehension. Journal of Educational Psychology, 82, 498-504.

Schraw, G., Dunkle, M. E. & Bendixen, L. D. (1995). Cognitive processes in well-defined and ıll defined problem solving. Applied Cognitive Psychology, 9, 523-538. 10 Nisan 2006'da http://web.ebscohost.com/ehost/pdf?vid=13&hid=105&sid=a63c1a96d15a-4695-aa35-62626af53d0f%40 sessionmgr109

Vandenberg, R.,J & Lance, C.E. (1998). A Summary of the issues underlying measurement equivalence and their implications for interpreting group differences. Research Methods Forum. 25 Ekim 2005'de http://www.aom.pace.edu/rmd/1998_forum_equiv_group_ differences.html

Vandenberg, R.J. & Lance, C.E. (2000). A Review and synthesis of the measurement ınvariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods, 3 (1), 4-70. 24 Ekim 2005'de http://orm.sagepub.com/cgi/ reprint/3/1/4.pdf

Wicherts, J. M. (2007). Group differences in ıntelligence test performance. Unpublished dissertation, University of Amsterdam. 16 Ocak 2008'de http://www.repository.naturalis.nl/document/ 44999

Gödelek, K. (2005). Güç iktidar bağlamında kadına yönelik şiddet. Muğla Üniversitesi Sosyal Bilimler Enstitüsü Dergisi (İLKE) G, 15, 97-107.

Lucke, J. F. (2005). The alpha and mean of Congeneric Test Theory: An extension of reliability and internal consistency to heterogeneous tests. Applied Psychological Measurements. 29 (1),65-81.