Bulletin of the *Transilvania* University of Braşov
Series VII: Social Sciences • Law • Vol. 6 (55) No. 2 - 2013

PROCEDURES FOR TRANSLATING AND EVALUATING EQUIVALENCE OF QUESTIONNAIRES FOR USE IN CROSS-CULTURAL STUDIES

A. ABUBAKAR[1,2] R. DIMITROVA[1] B. ADAMS[1,3]
V. JORDANOV[4] D. STEFENEL[5]

Abstract: *Taking into account the increased need to understand human behaviour across different cultural contexts there is a greater need than before to translate scales for use in large scale studies. Valid comparisons across cultures require that the translation process be accurate and that the scales being used measure the same psychological constructs across groups (i.e. measurement invariance). The current paper sets out to highlight the state- of-the-art procedures for translating scales and evaluating their invariance across cultural context. We first highlight the different ways in which item translation can go wrong and then discuss procedures for carrying out the translation process. Second, we discuss the construct of equivalence and then using data from more than 430 participants in 3 countries (Romania, United Kingdom and South Africa) we illustrate how one can evaluate measurement equivalence within a confirmatory factor analysis model. It is expected that the current paper will provide the reader with adequate background information on how to carry out translation of scales.*

Key words: *Test translations, Cross-cultural, Equivalance.*

1. Introduction

The current article sets out to highlight approaches to ensuring one adequately translated scales in a new cultural context. We first highlight procedures for translations, then look at the procedures for enhancing the quality of translation before finally describing statistical approaches that can be used to ensure that the translated scales are equivalent across contexts. Data from three cultural contexts, Romania, the United Kingdom (UK), and South Africa, are used to illustrate how one can evaluate measurement invariance.

There is an increased need to study human behaviour across cultural contexts. Consequently, there is a great need to

[1] Tilburg University, the Netherlands.
[2] Utrecht University, the Netherlands.
[3] University of Johannesburg, South Africa.
[4] National Sports Academy, Bulgaria.
[5] Contemporary Balkania, Greece.

translate scales that will be used in cross-cultural studies. Most of the standardized scales have been developed in English speaking countries. To use scales in non-English language speaking contexts, researchers are required to translate the English version into the language in which the scale is needed. When not carried out properly, translations could easily lead to poor reliability and validity in the data collected; consequently the value of an adequate translation cannot be overstated [1]. The literature is full of examples of cases where the study design and results are compromised due to problems arising from inadequate translations [2]. One example of where a translation can go wrong is when one uses a term that provides a clue of the required response in the target language. A classical example is this multiple choice item as cited by Van de vijver and Tanzer [3]. *Where do birds with webbed feet live? a) In the mountains, b) in the woods, c) in the sea, or d) in the desert.* The Swedish version of the scale asked where do birds with 'swimming feet' live making the item much easier to respond to for the Swedish speaking children. Translation problems may also arise due to differences in the grammatical and structural aspects of language. An example has to do with the differences in the structure of English when compared to other languages. In English, we do not need to include an article when mentioning a noun, however in other languages (e.g., Spanish, and Arabic's) an article is included, and this article carries important information.

During a translation process this differences in language can easily lead to serious difficulties. A good example of this problem is presented by Pena *et al* [4] while discussing issues related to the adaptation of the Peabody Picture Vocabulary Test–Revised (PPVT) [5] to the Spanish Test de Vocabulario en Imagenes Peabody [6]. PPVT is a single-word recognition task. The examiner tells the child a word and the child is expected to select one of four pictures that best depicts the given word. The original version (since was developed in English has no articles i.e., dog and not the dog); to ensure that 'equivalence' and avoid the situation where a child may receive an extra clue even the Spanish version requires that examiner says 'dog' not the 'the dog'. Yet naturally the Spanish in their day- to- day usage of language they would include an article in naming a dog. Pena et al. (2007) argue that '*Omitting the article could result in a functional difference unintentionally affecting test performance because Spanish-speaking children do not typically hear nouns without their articles'.*

A good translation aims to avoid all these potential pitfalls by producing a measure that is conceptually equivalent to the original one yet easy to read, to understand and use in the target language. To avoid the problems listed above and other potential bias, it is important to follow a systematic approach in translating and evaluating scales. Figure 1 below shows a step-by-step procedure of the translation and evaluation procedure.

2. Procedures for Translations

There are various approaches to translations including back-translations and committee approaches [7]. Each of these approaches has limitations and we recommend the use of an approach that combines these different approaches [8]. We recommend a procedure that comprises forward translation. During forward translation items from the original language are translated into the target language. To ensure the quality of the translation one has to carefully select translators. Preferably they should be bilinguals, who are familiar with the local context and the construct of interest [9].

Following translation by these bilinguals the back-translation can be requested from 2 independent bilinguals who are familiar with the language. The last step can be a harmonization stage where a panel of experts meet to discuss both the original items and the translated items. The panel will aim at identity faulty items, discussing alternative wording, and discuss the new wording until they reach a consensus. An important point to note is that this is an iterative process, where the three steps are repeated until one produces a scale that is conceptually equivalent to the original one and is easy to read in the target language.

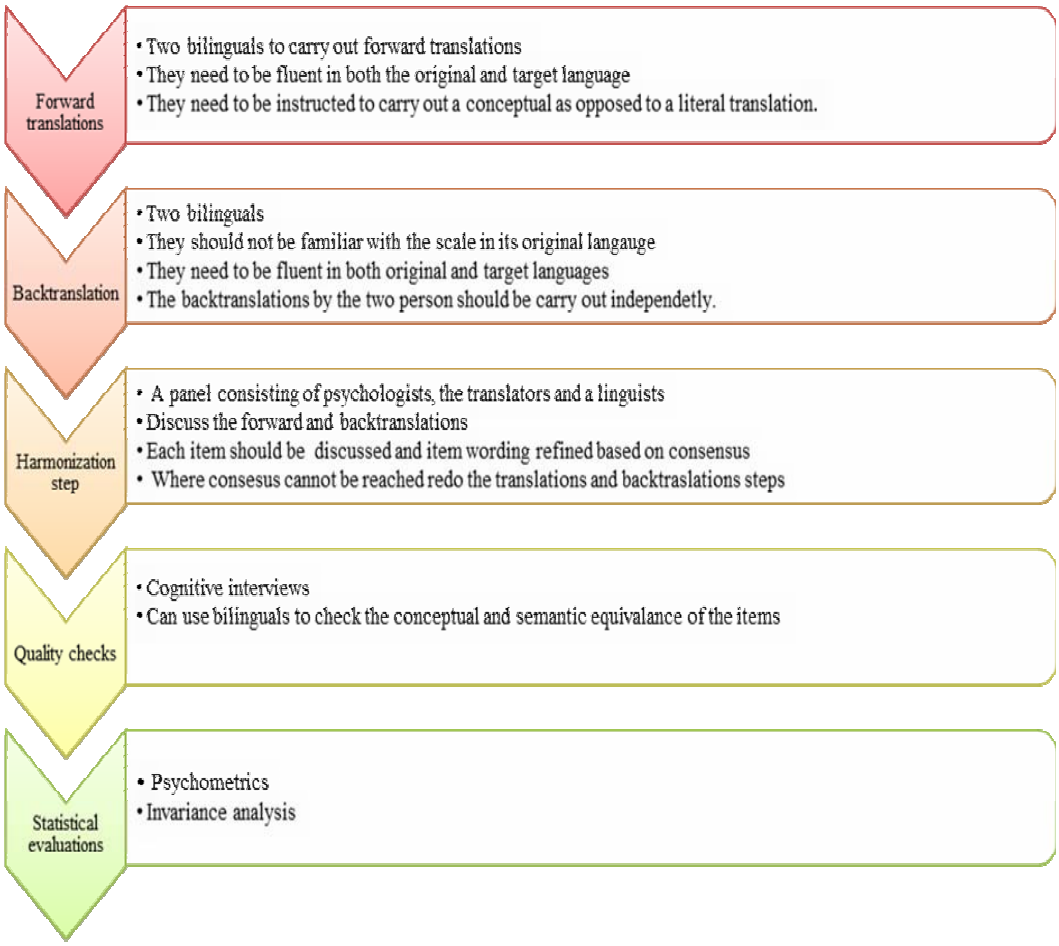

Fig. 1. *A summary of the Test Translation and Evaluation Procedures*
This figure was adapted from Abubakar and Van de Vijver [10]

3. Procedures for Evaluating and Enhancing the Quality of Translation

Having gone through the translation procedures, one cannot assume that the perfect scale has been identified. There are many strategies that can be used to evaluate the extent to which the translated items are of good quality. In this article we are going to discuss two such strategies which we think are not only easy to implement but also produce highly reliable evaluation of the scale.

3.1. Cognitive interviewing

This is a process aimed at investigating the cognitive processes by which a respondent answers to survey questions in particular as it relates to comprehension, recall, decisions and judgement [11]. The aim of this process is to detect problems in the questionnaire and correct those problems before the questionnaire is administered. There are two ways in which cognitive interviews can be carried out. First, we have the *Think aloud protocol*- in this approach the interviewer asks the person taking the test to say loudly their thoughts as they prepare to answer the question asked. In this way the interviewer hopes to understand the process by which the person comes up with the right answer. Second, we have the *Verbal probing technique*-in this technique the participants are asked about selected questions to examine the comprehension of the questions and highlight the reasons for their choices and to pick out problematic items. Questions used in this session need to include measures of comprehension, recall, and evaluation. Cognitive interviews basically involve small sample sizes (around 15) and it is also an interactive process where mistakes identified are corrected and the process repeated until one is satisfied with the measure.

3.2. The use of bilinguals

Bilinguals can be extremely useful in helping one investigate the degree to which the translated items are equivalent and identify any faulty items early enough. As a process bilinguals can be used in various ways to enhance the quality of the translation. Mallinckodt and Wang [12] present an elegant approach in which bilinguals were used to evaluate a translation of the Experiences in Close Relationships Scales [13] into Mandarin. In this study, a dual-language procedure was implemented where bilingual students were presented randomly with half of the scale in Mandarin and the other half in English. The English and Mandarin items were alternated. So each item on the scale was responded to by half of the students in English and half of the students in Mandarin. Analysis of each item was carried out to examine if there were items where students were more likely to fail when presented in a particular language. Such a procedure allows for the statistical evaluation of the extent to which items within a scale are differentially performed based on the language they were administered in. Items that show different patterns of performance based on the language of administration can then be reviewed and necessary changes implemented.

4. Statistical Procedures for Evaluating the Psychometric Properties of the Translated Scales

This section will be discussed using a practical example from the Satisfaction with Life Scale (SWLS) [14] which is a 5-item self-report instrument assessing subjective well-being. The scale is scored on a five-point Likert scale ranging from 1 (*Strongly disagree*) to 5 (*Strongly agree*), with higher scores indicating greater life satisfaction. Sample items include "I'm satisfied with my life" and "On the whole my life is next to my ideal". For the purpose of this illustration we used data from 438 participants from South Africa, Romania and UK A significant number of participants did not indicate their age, but for those who did, the mean age of the sample was 21.37 (SD = 3.06) with the South African sample being significantly younger (mean 19.68, SD = 1.61) compared to UK sample whose mean age

was 22.37 (*SD* = 4.14). The mean age of Romanians was 22.52 (*SD* = 2.76).

4.1. Psychometric Evaluation

An adequately translated scale is one whose basic psychometric characteristics are good. Two key psychometric characteristics are usually of interest. First, reliability which refers to the extent to which the scales provides consistent scores across items (internal consistency), time (test-retest reliability), and administrators (interrater reliability). We do not have the latter two forms of reliabilities but we did evaluate the internal consistency. In this case the alphas for the Romanian sample were excellent according to the rules of thumbs (alpha of .826), similar results were observed for UK, .820 while for South Africa it was slightly below the recommended value of .680; recommended value is .700 [15].

4.2. Evaluation of Measurement Invariance Using Confirmatory Factor Analysis

To evaluate the extent to which the scale measures the same construct in UK, Romania and South Africa one can carry out the four steps for evaluating measurement invariance through Confirmatory Factor Analysis (CFA). We highlight the steps that can be taken to evaluate the invariance of a scale within a CFA model.

1. First, to estimate the CFA of the model in each country. This step helps to check that the data fits in all countries as this is especially important when one has many countries and a single country with poor structure may significantly influence the observed results, leading to the whole dataset showing a lack of invariance. The quality of the fit needs to be decided based on set standards and suggestions from the literature. It is recommended that the following fit indices be assessed: Chi-Square (in an adequately fitting model the chi-square results should be non-significant. However, there is sufficient evidence to show that chi-square statistics are sensitive to sample size. Therefore if one has a large dataset chances are that the chi-square statistics will be significant. Consequently most people ignore this while making decisions on the adequacy of the specified model), the Tucker-Lewis Index (TLI < .95 excellent; < .90 good), the Comparative Fit Index (CFI < .95 excellent; < .90 good) [16] and Root Mean Square of Approximation (RMSEA > 0.06 excellent > 0.08 good). Testing a single factor model, indicated that the scale was unidimensional and that our data had a good fit to the hypothesized model. All the fit indices were above the recommended minimum standards. In addition, the factor loading were significant and substantial in both cases as shown in Figure 2 regarding details of the factor loadings. What these results indicate is that this scale has a unidimensional structure in all countries. Using only this set of findings one cannot justify comparing observed means in these two groups. To be able to compare group means one needs to carry out further analytic steps.

2. The next important step involves multigroup CFA where one will examine the configural model for all groups simultaneously. In this model, we tests if the structure of the scale is similar in all countries. No comparison is made between groups. If all the countries had good fitting models in step one, this step is expected to be ok. See table 2 for the fit indices from our analysis.

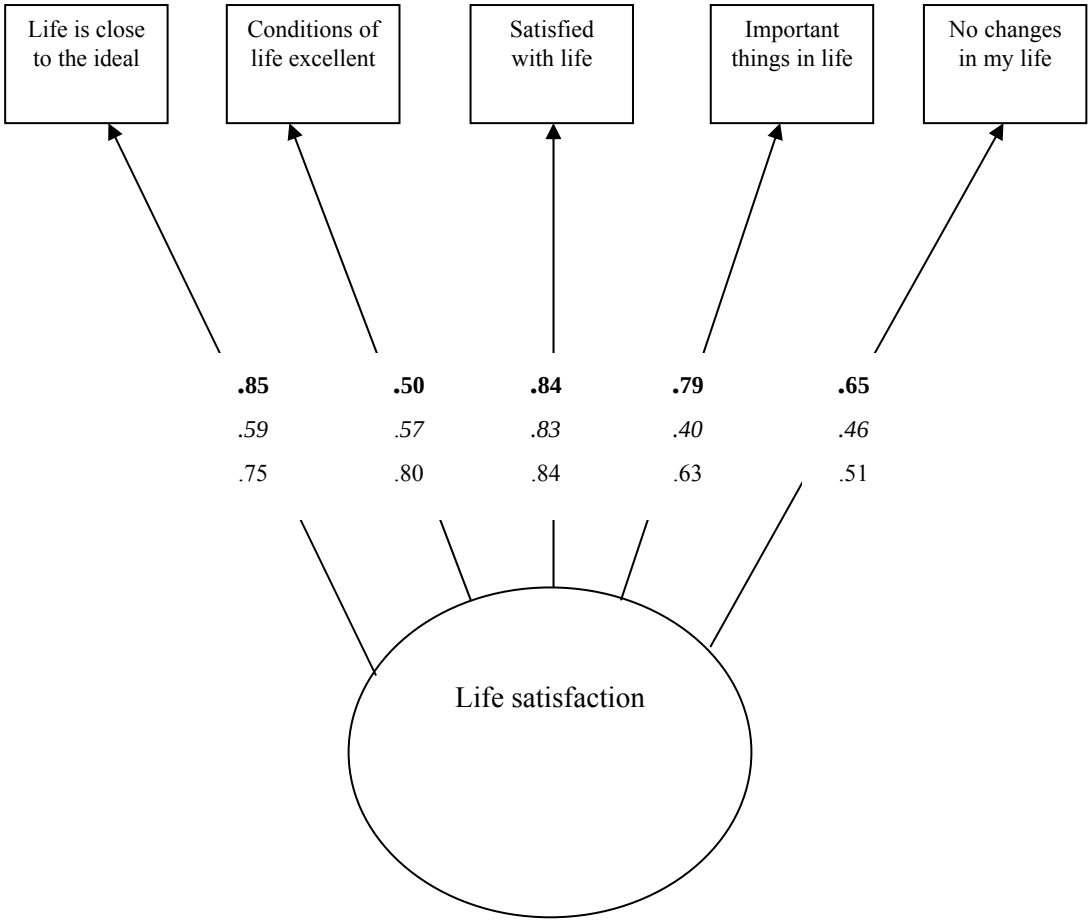

Fig. 2

3. The third step involves testing for metric invariance; here constrains are included in the analysis to evaluate the extent to which the factor loading would be equal across groups. In this step not only the fit indices have to be within the acceptable range, as discussed in step one but also additional criteria apply. Here one evaluates the quality of successive models. The delta CFI (the term used to refer to the change in CFI of less than .010 is considered desirable) and the smallest AIC figure represents a better fitting model [17]. Although our model has a good fit to the data the change in CFI was large (.021) which indicated that we did not yet achieve metric invariance. As suggested in the literature when one fails to meet these criteria, we need to identify the items causing misfit and free some factor loadings (partial metric invariance).

Results indicate that after releasing the loading on one item (item 2) we achieved a partial metric invariance.

Measurement invariance statistics for SWLS across three countries Table 1

Model	χ^2	df	χ^2 / df	RMSEA	TLI	CFI	ΔCFI	AIC
Unconstrained model	31.74	15	2.11	.051	.953	.976	-	121.74
Measurement weights	54.66	23	2.37	.056	.942	.955	.021	128.66
Partial Measurement weights	43.38	21	2.06	.050	.955	.968	.008	121.38
Measurement intercepts	99.75	33	3.02	.068	.914	.906	.039	153.75
Partial Measurement intercepts	52.78	23	2.29	.055	.945	.958	.010	126.78

The fourth step, involves testing for scalar invariance, which entails constraining the intercepts to be equal across groups and checking the fit for the model (in AMOS this is the measurement intercept model). If the model meets the rules of thumb as used for the metric invariance step i.e. (Step 3) you can assume that the observed means from both groups measure the same underlying construct. In our model this was not achieved. We therefore had to go an extra step and free some loadings, which enabled us to attain partial scalar invariance. We had to release 3 items, and the only invariant items were items 1 and item 4, while some people have argued that this is sufficient to compare observed means across groups others argue that in such a case the best approach would be to compare latent means. Further discussions on this can be read I several earlier publications e.g Milfont, & Fischer, (2010).

5. Conclusion

Adequate translations require a systematic and thorough approach that involves various iterative steps. The adequacy of translated materials needs to be evaluated using both qualitative (e.g. cognitive interviewing) and quantitative (e.g. Invariance analysis) approaches as illustrated in this paper. Approaching translations with vigour will contribute greatly to enhancing the validity of the cross-cultural studies.

References

1. Sireci, S.G., et al.: *Evaluating Guidelines For Test Adaptations A Methodological Analysis of Translation Quality.* In: Journal of Cross-Cultural Psychology, 2006. 37(5): 57-567.
2. Hambleton, R.K., L. Patsula, L.: *Adapting tests for use in multiple languages and cultures.* In: Social indicators research, 1998, 45(1-3): 153-171.
3. van de Vijver, F., Tanzer, N.K.: B*ias and equivalence in cross-cultural assessment: An overview*. In: Revue Européenne de Psychologie Appliquée / European Review of Applied Psychology, 2004, 54(2): 119-135.
4. Peña, E.D.: *Lost in Translation: Methodological Considerations in Cross‐Cultural Research.* In: Child Development, 2007, 78(4): 1255-1264.
5. Dunn, L.M., Dunn, L.M.: P*eabody picture vocabulary Test* (Revised). Circles Pine, MN. American Guidance Association, 1981.
6. Dunn, L., et al.: *Test de Vocabulario en Imagenes Peabody*. Circle Pines, MN. American Guidance Service, 1986.
7. Carlson, E.D.: *A Case Study in Translation Methodology Using the Health‐Promotion Lifestyle Profile* II. In: Public Health Nursing, 2000. 17(1): 61-70.

8. Holding, P., Abubakar, A., Wekulo-Kitsao, P.: A *Systematic Approach to Test and Questionnaire Adaptations in an African Context*. In: 3mc. 2008. Berlin, Germany: Available on: http://www.csdiworkshop.org/pdf/3mc2008_proceedings/session_09/Holding_Abubakar_oct.pdf. Accessed: 11/2013
9. Hambleton, R.K., Kanjee, A.: I*ncreasing the validity of cross-cultural assessments: Use of improved methods for test adaptations.* In: European Journal of Psychological Assessment, 1995, 11(3): 147.
10. Abubakar, A., Van de Vijver, F.J.R.: *How to adapt tests for use in Sub-Saharan African*. In: *Applied Developmental Psychology Perspectives from Africa*, A. Abubakar and F.J.R. Van De Vijver (Eds.). Submitted, Springer Publishers, New York.
11. Willis, G.B.: *Cognitive interviewing: A tool for improving questionnaire design*. Thousand Oaks. Sage, 2005.
12. Mallinckrodt, B., Wang, C.-C.: *Quantitative Methods for Verifying Semantic Equivalence of Translated Research Instruments: A Chinese Version of the Experiences in Close Relationships Scale*. In: Journal of Counseling Psychology, 2004, 51(3): 368.
13. Wei, M., et al.: *The Experiences in Close Relationship Scale (ECR)-short form: Reliability, validity, and factor structure.* In: Journal of personality assessment, 2007, 88(2): 187-204.
14. Diener, E., et al.: *The satisfaction with life scale*. In: Journal of personality assessment, 1985, 49(1): 71-75.
15. Cicchetti, D.: *Guidelines, criteria and rules of thumbs for evaluating normed and standardized assessment instruments in Psychology*. In: Psychological Assessment, 1994, 6: 284-290.
16. Hu, L., Benter, P.: *Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives.* In: Structural Equation Modelling, 1999, 6: 1-55.
17. Milfont, T.L., Fischer, R.: *Testing measurement invariance across groups: Applications in cross-cultural research*. In: International Journal of psychological research, 2010, 3(1): 111-130.