Cross-cultural adaptation and psychometric validation of research instruments: A methodological review

S. M. Yasir Arafat¹, Hafizur Rahman Chowdhury¹, M. M. A. Shalahuddin Qusar¹, M. A. Hafez²

ABSTRACT

Background: Psychometrics has a very important role and with the increase in the number of cross-cultural research projects, there is need to adapt the scales to measure health status. Objective: It was aimed to describe the validation process comprehensively based on recent practices and to overview a practical guideline of the validation study. Materials and Methods: Search was done in PubMed, PubMed Central, and Google Scholar with searching keywords ranging date from January 2000 to December 2015. Finally, 78 articles were selected for review and thorough methodological review was done in accordance with the review objectives. Result: Review revealed that cultural adaptation followed similar steps with few variations. There was huge variation in study design, sampling technique, and sample size. Cronbach’s alpha and test-retest reliability were mostly used forms of reliability with variations in assessment of validity. Exploratory factor analysis was used more for statistical analysis and followed similar criteria with few variations. Conclusion: Cross-cultural measurement of health outcomes is really an important and challenging issue that needs elaborative guideline to take appropriate steps to solve problems.

KEY WORDS: Cross-cultural adaptation, psychometric validation, review article, scale validation, validation methodology

INTRODUCTION

Psychometrics has very important role in public health, psychiatry, primary health care and many other fields even in health promotional strategy for measuring the attitude [1]. With the increase in the number of multinational and multicultural research projects, the need to adapt health status measures for use in other than the source language has also grown rapidly [2]. When using self-report measures, it is important to realize that they are potentially vulnerable to distortion due to a range of factors including social desirability, dissimulation and response style [3]. Consequently, there is much emphasis on using standardized and validated research instruments to measure the responses [4]. As a result, there needs to adapt the instruments accordingly to measure the responses appropriately. Subsequently, culturally adapting has many advantages over developing a new one. For example, it reduces the costs and the time spent in development and allows for using the instrument, which has been widely used before, to make intercultural comparisons [5]. The cross-cultural adaptation of a health status self-administered questionnaire for use in a new country, culture, and/or language necessitates use of a unique method, to reach equivalence between the source and target versions of the questionnaire [2]. This includes translation, adaptation, the assessment of validity (content validity, face validity, construct validity, and criterion validity), reliability (repeatability and internal consistency) and responsiveness [6]. Translation of the instrument is a crucial step in the validation process [7], but the terms “adaptation” and “translation” are distinct, and the former has been used most often because it includes all the processes concerning the cultural fit of the instrument beyond mere translation. Translation is merely the first stage of the adaptation process. When adapting an instrument; cultural, idiomatic, linguistic, and contextual aspects concerning its translation should be considered [8]. The process of transcultural adaptation involves the development of versions of an assessment instrument that are equivalent to the original, but at the same time, linguistically and culturally adapted to a different context than the original. Therefore, the adaptation of assessment instruments allows comparisons among results of investigations conducted in different cultures, aiding the exchange of information within the international scientific community, and decreasing costs and time spent in the process [9]. Reliability of a measure refers to the ability of a questionnaire to determine that a measurement yields reproducible and consistent results [6,10,11]. Reliability analysis included internal consistency reliability, split-half reliability, test-retest...
Face validity is the ability of an instrument to be understandable and relevant for the targeted population. It concerns the critical review of an instrument after it has been constructed and generally includes a pilot testing [6, 10].

**Property** | **Definition**
---|---
Content validity | The ability of an instrument to reflect the domain of interest and the conceptual definition of a construct [6, 10, 15].
Face validity | The ability of an instrument to be understandable and relevant for the targeted population. It concerns the critical review of an instrument after it has been constructed and generally includes a pilot testing [6, 10].
Construct validity | The extent to which a new measure is related to specified variables in accordance with an established theory or hypothetical construct [6, 11, 14, 15].
Convergent validity | The degree to which scores on a measure associate with scores on other measures that are intended to assess similar constructs [6, 16].
Divergent validity | Involves that item within any one subscale should not correlate too highly with external items or with the total sum-score of another subscale [6].
Discriminative validity | The ability of an instrument to distinguish between groups that are expected to differ based on their clinical diagnosis or other characteristics [16].
Known group validity | The ability of an instrument to be sensitive to differences between groups of patients that may be anticipated to score differently in the predicted direction [6].
Criterion validity | The assessment of an instrument against the true value, or a standard accepted as the true value. It can be divided into concurrent validity and predictive validity [6].
Concurrent validity | The association of an instrument with accepted standards [6].
Predictive validity | The ability of an instrument to predict future health status or test results. Future health status is considered as a better indicator than the true value or a standard [6].
Internal consistency | The ability of an instrument to have interrelated items [6, 10].
Test-retest reliability | The ability of the scores of an instrument to be reproducible if it is used on the same patient while the patient’s condition has not changed (measurements repeated over time) [6, 10, 13].
Responsiveness | The ability of an instrument to detect change when a patient’s health status improves or deteriorates [6, 14].
Floor/ceiling effects | Descriptive statistics were calculated to identify floor/ceiling effects, which were considered to be present. When >15% of the subjects obtained the lowest or highest possible scores, indicating the proportion for whom, respectively, no meaningful deterioration or improvement in their condition could be detected since they are already at the extreme of the range [14].

**MATERIALS AND METHODS**

Search was done in PubMed, PubMed Central and Google Scholar with searching keywords ranging date from January 2000 to December 2015. After completion of download, exclusion of repetition was done and then articles were chosen purposively according to the inclusion and exclusion criteria. From total 2529 articles, finally 78 articles were selected for review.

**Inclusion Criteria**

a. Full downloadable articles in pdf form
b. Article available only in English language.

**Exclusion Criteria**

a. Multiple validation studies by same principal author on different scales having same parameters
b. Validation studies using only CFA
c. Validation studies other than behavior measure scales.

data. Total article downloads 2529
b. After exclusion of repetition (n = 484), total articles were 2045
c. From 2045 articles, 547 article were chosen purposively by assessing title and abstracts and chosen based on proportionally more in recent years
d. 547 articles screened, 174 were chosen for review after assessing the full text
e. Finally, 78 articles were chosen after applying the exclusion criteria [Figure 1].
RESULTS

Data Characteristics

Among the articles, there were 64 original articles, 12 review articles, and 2 other types (invited article and short communication). Yearly distribution of article was 2000-2; 2001-1; 2002-1; 2003-1; 2005-3; 2007-1; 2008-1; 2009-3; 2010-4; 2011-6; 2012-6; 2013-11; 2014-23; 2015-15.

Cross-cultural Adaptation

For cross-cultural adaptation, there are certain changes suggested in the literature. Health measure outcome responses may vary on the basis of country, culture, and language [Table 2]. For these changes, there needs to follow the proper translation process [Figure 2] along with adequate changes to ensure the equivalences. Semantic equivalence: Ensures the equivalence of meaning as the translated version needs to mean the same with the original. Idiomatic equivalence: Colloquialisms, or idioms, are difficult to translate. The expert committee has to ensure considering the original and the target culture for equivalent expression. Experiential equivalence: Items are seeking to capture and experience of daily life; however, often in a different countries or cultures, a given task may simply not be experienced (even if it is translatable). The questionnaire item would have to be replaced by a similar item that is in fact experienced in the target culture. Conceptual equivalence: Often words hold different conceptual meanings between cultures that need to be addressed.

Study Design

Among the articles, the study design was not mentioned clearly in majority of the articles. Those who mentioned clearly, mentioned the validation study as cross-sectional study design, few of the articles mentioned as methodological study.
Sampling Technique and Sample Size

Among the studies, both probability and non-probability sampling techniques were used as a sampling technique having more deviation to non-probability sampling. There are variations in the sample size based on the item number, sample size estimation formula as well as the authors’ choice. Their lacks any specific guidelines as a result different authors used different formula as well as different sample size. Some authors estimated sample size on the basis of factor analysis (EFA), some estimated using the regression formula of sample size estimation while more authors chose the item sample ratio method to estimate the sample size. Regarding EFA basis of sample size estimation, recommendations to ensure the sample size ranges from 100 to 250 while other recommendations mentioned to ensure >300 to ensure the good statistical estimation. Regarding sample size estimated by regression formula, there are variations based on researchers’ criteria. The item sample ratio sample size estimation process lacks any definitive ratio. Researchers use minimum 2 to maximum 20 people per item to estimate the sample size that is assumed arbitrary. Some authors use Kaiser–Meyer-Olkin (KMO) sampling adequacy test to ensure the adequate sample size. Different authors took different KMO values as the level of adequacy that ranges from 0.50 to 0.60.

Reliability Assessment

Reliability was assessed in forms of internal consistency, test-retest reliability, inter-rater reliability, and split-half reliability. Almost every article assessed internal consistency form of reliability and in all article, it is measured by Cronbach’s alpha having level of ≥0.70. The majority of the articles assessed test-retest reliability along with internal consistency and that was assessed by different statistical measures such as Wilcoxon non-parametric statistical test, intra-class correlation coefficient (ICC), Pearson’s correlation coefficient, Spearman correlation coefficient, Kappa coefficients, paired t-tests, and Lin’s

Table 2: Areas where adaptation is recommended for cross-cultural validation of instruments [2]

<table>
<thead>
<tr>
<th>Wanting to use a questionnaire in a new population described as follows</th>
<th>Results in a change in</th>
<th>Adaptation required</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Culture</td>
<td>Language</td>
</tr>
<tr>
<td>A</td>
<td>Use in same population. No change in culture, language, or country from source</td>
<td>-</td>
</tr>
<tr>
<td>B</td>
<td>Use in established immigrants in source country</td>
<td>✓</td>
</tr>
<tr>
<td>C</td>
<td>Use in other country, same language</td>
<td>✓</td>
</tr>
<tr>
<td>D</td>
<td>Use in new immigrants, not English-speaking, but in same source country</td>
<td>✓</td>
</tr>
<tr>
<td>E</td>
<td>Use in another country and another language</td>
<td>✓</td>
</tr>
</tbody>
</table>

Figure 2: Graphical representation of translation process for validation study [2]
correlation coefficient. Few of the articles assessed inter-rater reliability by Kappa statistics; Student’s t-test statistical analysis and split-half reliability were assessed by Pearson’s correlation coefficients between two halves of the items.

Validity Assessment

Validity is assessed in forms of face validity, content validity, criterion validity, convergent validity, construct validity, divergent validity, concurrent validity, predictive validity, and internal validity. The validity assessment is much customized on the basis of researchers’ criteria, instrument feasibility and statistical analysis. There are variations in assessment and way of measurement.

Statistical Analysis and Factor Analysis

Statistical analysis is done to assess structural validity, sampling adequacy, internal consistency, factor rotation, factor identification, item retention, and other steps of analysis that options are predesigned in the available software. Factor analysis was done to assess the structure of the construct. Among the two major classes of factor analysis: EFA and CFA; EFA in the form of principal component analysis with Varimax rotation was by far the most common choice. Factors retention was done based on criteria of eigenvalue ≥1 and the scree plot, the graphical representation of eigenvalues. Item reduction was done by assessing the loadings after the factor rotation having value ≥0.30 with variations such as ≥0.32 and ≥0.40 based on researchers’ criteria.

DISCUSSION

This literature review aimed to overview common validation guidelines based on recent practices and focused on 78 psychometric validation studies published between January 2000 and December 2015 with more emphasis on recent articles during the data extraction process.

Cross-cultural Adaptation

Although most transcultural adaptation studies follow similar methodologies, there is no consensus in the literature regarding how the process should be conducted [9]. Most articles follow the process based on Beaton et al. [2] having steps of Initial translation, synthesis of the translations, back translation, expert committee review (semantic equivalence, idiomatic equivalence, experiential equivalence, and conceptual equivalence), test of the prefinal version and final questionnaire with some changes according to the study condition [4,5,7-9,12,13,17-40]. This process [Figure 2] ensures the best-guided translation and adaptation process fetching the consideration of standard translation, adequate equivalence and furthermore changes supported by the pretesting. Although the studies follow the more or less similar structure, there is no absolute guideline that must be followed. However, guideline described by Beaton et al. [2] is the mostly used and practiced guideline.

Study Design

The study design was not mentioned directly, and there is some hesitation regarding the matter. Those who mentioned, in the majority of the articles, it was mentioned as cross-sectional study design [12,19,26,41,42] and only a few of them were mentioned as a methodological study [9,20] that needs attention regarding the study design of a validation study. The review seems that the study design of a validation study is in the gray area, and there is much space to make further explorations regarding the study design. Authors intended to ascertain the study design as a methodological study or as itself a separate entity because of having separate scientific methods of a validation study that is distinct from cross-sectional or other study design.

Sampling Technique and Sample Size

Most articles have no justification regarding sample size, sampling technique, and sampling formula for sample size estimation. Sampling techniques were followed with having more interest in non-probability sampling [14,28,35,41,43,44] without having any justification. There is no consensus to define sample size with the same rigor as found in most clinical research based on clinical or biological criteria (e.g. arbitrarily determined sample size or subject to item ratio) [6,45]. Based on EFA sample size estimation recommendations to ensure an absolute minimum of 100-250 subjects, while other recommendations mentioned the following guidance: 100 = poor, 200 = fair, 500 = good, 500 = very good, and ≥1000 = excellent [6,21,46]. According to the EFA based estimation, sample size is mostly estimated arbitrarily. Some authors estimated sample size by following the regression formula which makes variations in sample size based on researchers’ criteria [17,20,23,39,46-49]. Majority of the author used item sample ratio method for sample size estimation with an arbitrary margin of 2-20 [6,9,10,35,41,44,45,46,50]. Previous reviews showed that subject to item ratios of ≤10:1 covers 63.2% studies [50]. Some authors used KMO sampling adequacy during statistical analysis for sample size adequacy [1,10,18,19,21,27,32,33,40,41,46,51-56] as routine part of EFA. The level of KMO for adequacy was interpreted differently among the authors mentioned as 0.6 [41,52,54] and 0.5 [46,53]. Hence, there are variations in regards to the sample size estimation, sampling technique, and sample size estimation formula. Researchers exercise the freedom in this aspect of the validation study in accordance to the appropriate situation and explanation.

Reliability Assessment

Among the different reliability forms, internal consistency form of reliability was measured by Cronbach’s alpha having level of ≥0.70 with few variations. The recommended level of Cronbach’s alpha is <0.50 unacceptable, 0.50-0.59 poor, 0.60-0.69 questionable, 0.70-0.79 acceptable, 0.80-0.89 good, and ≥0.90 excellent [7,11,29,40,43,44]. Cronbach’s alpha is one of the psychometric indicators most commonly used to determine
the reliability and internal validity of an instrument [21]. There are variations among the researchers to use a single level of the Cronbach’s alpha as the level of significance. Second form of reliability; test-retest reliability was assessed by Wilcoxon non-parametric statistical test [10,19,20], ICC [9,12-14,20-27-29,34,37,39,40,43,48,55,57-59], Pearson’s correlation coefficient [13,41], Spearman correlation coefficient [60], Kappa coefficients [56], paired t-tests [9], and Lin’s correlation coefficient [61]. In assessing this form, there is also much freedom exercise probably due to variations in culture, outcome, and situations. The inter-rater reliability was measured by Kappa statistics [22,29], Student’s t-test statistical analysis [5], and Split-half reliability was assessed by Pearson’s correlation coefficients [12] between two halves of the items. The acceptable ICC level varies as acceptable considered when ICC is ≥0.70 [40], >0.60 [29], and ≥0.65 [48]. Other recommendation regarding ICC value is 0.00-0.10 virtually none, 0.11-0.40 slight, 0.41-0.60 fair, 0.61-0.80 moderate, and 0.81-1.0 substantial [43]. Recommendation regarding Kappa statistic is ≤0.20 slight, 0.21 to <0.40 fair, 0.40 to <0.60 moderate, 0.60 to <0.80 substantial, and 0.81-1.00 excellent [43]. This form also has variations in regards to the universalization of the parameter as well as fixed significance level.

Validity Assessment

Different forms of validity were assessed differently.

Face validity

It is the easiest and weakest form of validity and it was assessed by ensuring standard back-translation process [2,27], by critical review and expert panel opinion [6,13,48], feasibility, readability, consistency of style and formatting, and the clarity of the language used [10].

Content validity

Content validity was measured as well as ensured by ensuring standard back-translation process [2,27], by literature review and expert panel opinion [6,10,13,25,48], and by experts with content validity index [49,53,62]. In different studies and situations, researchers used different options in accordance to the feasibility of the study domain.

Construct validity

Assessed by statistical analysis done by factor analysis [10-12,18,41,48,49,51,55,56,59,62], by comparing with other similar instruments [17], by comparing with other similar instruments with Spearman’s correlation coefficient [21], using Pearson’s linear correlation with comparison of similar instruments [13]. Based on the basic concept of construct validity [Table 1], it was assessed by comparing the instrument with other instruments with the help of the statistical methods which demonstrated variations in respect of statistical method that are already mentioned. Researchers here again practiced the freedom of the situation and condition based on accepted scientific method.

Criterion validity

Assessed by comparison with the gold standard instrument [6,15,48,59]. Like all other forms of validity, there were also variations in the assessment of criterion validity. Basic concept of criterion validity [Table 1] demands the comparison with the gold standard instruments. However, in case of validation, many of the cultures lack other validated gold standard one which made the researchers a troublesome situation as well as creates a field of variations.

Convergent validity

To ascertain the convergent validity, authors used Pearson’s correlation coefficient with comparison of other similar instruments [23,41] and Spearman correlation coefficient [22] with comparison of other similar instruments [40,51] as the validity demands similar statistics [Table 1]. The different statistical processes used based on the data characteristics, nature of the measures, and feasibility in regards to the study objectives.

Concurrent validity

Concurrent validity was measured by comparing the scale with another one by correlating with Pearson correlation coefficient [9] as it demands such assessment based on the concept of concurrent validity [Table 1].

Internal validity

The internal validity (consistency) was assessed using Cronbach’s alpha coefficient of internal consistency [27]. Validity assessment reveals diversions on the basis of researchers’ criteria, instrument feasibility, and statistical analysis in the line of the validity concept as well as variations in ways of assessment.

Statistical Analysis and Factor Analysis

There are two major classes of factor analysis: EFA and CFA. In EFA, the investigator has no expectations of the number or nature of the variables, and as the title suggests, is exploratory in nature. That is, it allows the researcher to explore the main dimensions to generate a theory or model from a relatively large set of latent constructs often represented by a set of items. Whereas in CFA the researcher uses this approach to test a proposed theory [33].

Factor retention

To ascertain the number of factors in a measure, the criteria used based on having eigenvalue ≥1 and the scree plot [10,17,18,23,27,29,33,39,40,46,47,50,53-58,63]. Factors that were not full filling the criteria were dropped from the measure. Authors used the criteria much routinely as well as strictly.

Item reduction

In consideration of item reduction, authors used factor rotation statistics and retained the items with loadings after the factor...
rotation having value $\geq 0.30$ [23,44,55,56,62,64] with variations such as $\geq 0.32$ [27,50] and $\geq 0.40$ [33,39,47,53,65] based on researchers' criteria. In this process, authors took different values with the same statistics with their preference.

CONCLUSION

Cross-cultural measurement of health outcomes is really an important as well as a challenging issue that needs proper attention and elaborative guideline to get the real scenario and to take appropriate steps based on it. Overall literature review revealed there are variations in steps of the validation study in different aspects. This review structures and recommends a comprehensive guideline to adapt and validate the health outcome measures in different cultures. Although it is a comprehensive review regarding the steps of a validation study, articles in other languages, premium articles, other database searches, and articles published in longer duration might contribute more. Further extensive systematic reviews can be conducted to explore and to modify the guideline.

Authors’ Concern

Authors raise the concern regarding the study design of a validation study. Authors recommend to use validation study as a separate study design instead of cross-sectional as every validation study. Authors recommend to use validation study as a separate study design instead of cross-sectional as every validation study. Cross-cultural measurement of health outcomes is really an important as well as a challenging issue that needs proper attention and elaborative guideline to get the real scenario and to take appropriate steps based on it. Overall literature review revealed there are variations in steps of the validation study in different aspects. This review structures and recommends a comprehensive guideline to adapt and validate the health outcome measures in different cultures. Although it is a comprehensive review regarding the steps of a validation study, articles in other languages, premium articles, other database searches, and articles published in longer duration might contribute more. Further extensive systematic reviews can be conducted to explore and to modify the guideline.

Authors’ Concern

Authors raise the concern regarding the study design of a validation study. Authors recommend to use validation study as a separate study design instead of cross-sectional as every validation study.

ACKNOWLEDGMENTS

Authors thank Monon Psychiatric Hospital for logistic support.

REFERENCES


© SAGEYA. This is an open access article licensed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted, noncommercial use, distribution and reproduction in any medium, provided the work is properly cited.

Source of Support: Nil, Conflict of Interest: None declared.

Arafat, et al.: Psychometric validation