Translation or Development of a Rating Scale: Plenty of Science, a Bit of Art

Perhaps one of the greatest challenges in psychiatric research is the selection of an appropriate scale or measure to examine psychological constructs of interest. For a keen researcher, several questions at once spring to mind. What construct(s) does it measure? Does it have sound psychometric properties (reliability and validity)? Is it a screening or diagnostic tool? Is it valid for our setting and culture? Is it norm-referenced or criterion-referenced? Can the cut-offs suggested by the author be directly applied to our setting? Specifically, these questions assume relevance as there are several measures available for evaluating each construct.

Sometimes, a recent systematic review of available measures makes the job easier to choose a scale with good psychometric properties. However, the scale might not be available in regional languages, which is a necessity if it is a self-rated tool. In such situations, ‘translation’ of the tool using a standard protocol (e.g., the forward-translation and back-translation method of the World Health Organization [WHO])[3] may be required (Box 1). Also, some items of a given scale may not be culturally relevant. Consequently, some form of ‘adaptation’ of the items or format may be needed. At times, the need for adaptation only becomes apparent while translating a scale. Hence, ‘translation and adaptation’ of a scale is commonly used together. Furthermore, the psychometric properties of the translated or adapted scales may not be known and are required to be examined before they can be used in research, specifically during cross-cultural studies. More rarely, if the available scales do not meet the requirements of research or the constructs under investigation are not adequately captured by extant scales, there may be a need to develop a new psychological scale (Box 2). The steps involved in this process are different from that of translation and adaptation of an existing scale or measure (Box 3).

Translation and adaptation of existing scale
In this issue, Grover and Dua[2] have translated, adapted, and validated three scales on religiosity into the Hindi language. It is not uncommon to see simple translated versions of English language tools, without any validation, being used in research. These authors have used a standard method of translation by bilingual experts, followed by expert panel evaluation, then pretesting on 20 participants, and finally back-translation by independent bilingual experts. Although this sequence is somewhat different from the guidelines put forth by WHO, it follows the four standard techniques suggested by Brislin[3] to maintain equivalence between the original and translated measure: a) back-translation method, b) bilingual technique, c) committee or expert team approach and d) pretest procedure. Indeed, current approaches to translation have evolved from the Brislin’s classic back-translation technique to include more bilingual and bicultural experts to generate a consensus on the final version.

The scales were administered to 132 respondents in the first round. This was followed by a second round of application of the same instruments, after 3-7 days, in either Hindi (n = 61) or English (n = 71). Given the short turnaround time between testing and retesting, memory effects cannot be ruled out. Also, as the same participants were used for all three scales, there is a possibility of respondent fatigue, which could have affected the results.

There was high correlation between the itemized scores on the Hindi and English versions of all three scales. This was indicative of good cross-language equivalence of these tools, which makes them suitable for use not only in the Indian setting but also for comparison with global studies. There are higher chances of type I error in this approach. However, p values less than 0.001 were found for all comparisons, indicating that such a risk is low in this case. Test-retest reliability, internal consistency, and split-half reliability of the Hindi version for all three scales were in the acceptable range. An interrater reliability exercise could have been a meaningful addition. Further, the sample predominantly comprised of urban residents with a minimum of ten years of formal education, limiting the generalizability.
Scale development
Singh et al.[4] developed the Comprehensive Satisfaction Index (ComSI) to assess levels of well-being, happiness, and life satisfaction among elderly individuals and examined its psychometric properties. The need for the scale was justified by citing the unique health demands in the elderly and the changing family milieu that has compromised the safety and security of this group. Notably, there is little theoretical discussion to support a strong conceptual foundation for the scale and the putative domain content. This assumes importance because a top-down or deductive process (where the review of literature serves as the guide for item generation) appears to have been followed for domain identification, and this approach has to be based on robust theoretical foundations.[5] Instead, the domain identification was done by a panel of three experts, and item generation was performed by three independent experts.

Subsequently, 5 independent experts performed a content validation exercise that led to the development of a 26-item tool. However, the details of the item pool generated and the number of the items that had to be removed are not stated. Furthermore, some items were modified and shaped following pretesting in 30 individuals and expert consensus. Whether these 30 individuals were drawn from the same target population of interest and whether they were part of the larger sample used in the same study (n = 260) are unclear. The assertion that the “feasibility” of the scale was assessed through the pilot sample could possibly mean validity, as pretesting is part of the process of establishing content validity.

Applying an item-response ratio of 1:10, responses of 260 rural subjects were obtained. Hence the scale may not be applicable to urban respondents in whom the underlying needs and issues are likely to be different. Apart from age, gender, and socio-economic status, other sample characteristics are not presented, thereby raising concerns regarding sample representativeness and generalizability.

Dimensionality of ComSI was assessed using principal component analysis,[6] a procedure similar to exploratory factor analysis. The sample was adequate for factor analysis; however, an inter-item correlation matrix (which allows identification of items that correlate poorly and thus be a source of error and unreliability)[7] was not performed. To identify the number of factors, only Kaiser criterion[8] of eigenvalue >1 was used. Using multiple methods (e.g., scree plot or Horn’s parallel analysis)[9] could have increased the reliability.

Notably, item no 17 did not load onto any of the factors, indicating that the item may either not apply to the setting or need to be reworded as it was understood incorrectly by the respondents. Also, reliability statistics (internal consistency), a necessary pre-condition for validity, was not reported. Convergent validity was checked with the WHO Quality of Life (QOL) Scale, but discriminant and criterion validity were not examined.

Bedi and Varma[10] developed a new scale, Positive Temperament Inventory (PTI), to tap the positive temperamental attributes of Indian adults. There is a persuasive account of the relative neglect of positive emotionality in the extant literature. This and the unsuitability of Western instruments in Indian setting were key drivers for developing the scale. Domain identification and item generation were done through a deductive process[5] following a literature search, identifying items that represented the most relevant and common behaviours in the Indian setting and rewording them for cultural compatibility.

However, the semantics of the paper leaves room for some confusion. It is mentioned that “15 factors” were reduced to “6 factors” containing 6 items each (making it a 36-item scale), based on the opinion of 2 experts, because these 6 factors had more putative items related to positive temperament. However, this approach may have potentially excluded less frequent temperamental attributes.

Subsequently, exploratory and confirmatory factor analyses were carried out, separately, in two different but demographically similar samples. Convergent validity was assessed through the prosocial domain of the Strength and Difficulties Questionnaire (SDQ), while divergent validity was assessed through the correlations of factors with the neuroticism subscale of International Personality Item Pool (IPIP). However, the choice of SDQ to check validity is not appropriate because the age group intended to be covered by the SDQ (3-16) and the PTI (18-80) do not overlap.

The results of the two-stage factor analysis with sample adequacy, the goodness of fit statistics, and cut-offs for factor loading are explained well. Observed convergence between Kaiser criteria and scree plot, as well as good reliability statistics, adds certainty to the results. How two second-order factors (temperamental positivity and dynamic positivity) were arrived at using the correlation matrix of the four first-order factors was also explained adequately. However, the
Menon and Praharaj: Development of rating scales

Box 1: World Health Organization (WHO) protocol for scale translation
1. Forward translation: Translation to the target language by a bilingual individual aiming for conceptual equivalence of the items
2. Expert panel: A bilingual expert panel identifies and resolves the discrepancies between the two versions.
3. Back-translation: Translation back to original language by an independent person not familiar with the measure
4. Pretesting: On target participants followed by in-depth interviews or focused groups to identify any unacceptable words or expressions.

Box 2: When to develop a new scale or measure?
1. When existing scales do not measure the intended construct or do not adequately cover key attributes of the construct of interest
2. When the existing scales do not address the needs of the intended target population, which may differ psychologically from the original population in whom the scale was validated
3. When existing scales have become outdated due to cultural, societal, or generational change, and there is a need for new measures that may better capture topical issues
4. When the adapted or translated version does not have psychometric or dimensional stability across culturally and linguistically different groups
5. When novel research hypotheses, involving new constructs, are propounded.

Box 3: Steps involved in scale development
1. Define the construct and the context: Based on literature review, consultation with subject experts and a sound theoretical foundation
2. Selecting response format: The scale format (e.g., Likert type) and number of response options are selected
3. Item pool generation: Generation of items that adequately sample the content domains of interest
4. Item evaluation: A panel of experts examine the quality and relevance of the items, which may be modified or refined based on their input
5. Pilot testing: This is done through an exploratory study in order to create a more parsimonious representation of the original items
6. Testing for psychometric properties: Reliability assessment includes internal consistency, split-half reliability, and test-retest reliability. The dimensionality of the scale is examined through exploratory and confirmatory factor analysis. Other validity measures include convergent, discriminant, and criterion validity.

omission of findings of the inter-item correlation matrix limits the understanding of prerequisites of factor analysis. This is particularly relevant because moderate correlations were noted between the first-order factors. Also, the inclusion of a self-selected sample points to a possible selection bias, which is not acknowledged in the paper.

To sum up, using culturally validated measures lends greater credibility to any research. From this perspective, it is gladdening to note the increasing research focus on developing and validating culturally compatible instruments. This will, no doubt, enhance good quality data from our settings and facilitate comparisons with global literature. The readers are advised to read the three articles discussed here thoroughly along with this editorial in order to understand the nuances involved in the development of a psychological measure, which can be aptly described as plenty of science and a bit of art.

Vikas Menon, Samir Kumar Praharaj

Department of Psychiatry, Jawaharlal Institute of Post Graduate Medical Education and Research, Dhanvantri Nagar, Puducherry, 1Department of Psychiatry, Kasturba Medical College, Manipal, Karnataka, India

Address for correspondence: Dr. Vikas Menon
Department of Psychiatry, JIPMER, Puducherry - 605 006, India.
E-mail: drvmenon@gmail.com

REFERENCES
1. WHO | process of translation and adaptation of instruments [Internet]. WHO; Available from: https://www.who.int/substance Abuse/research_tools/translation/en/.[Last cited on 2019 Oct 12].
2. Grover S, Dua D. Translation and adaptation into Hindi of central religiosity scale, brief religious coping scale (Brief RCOPE), and Duke University Religion Index (DUREL). Indian J Psychol Med 2019:6;556-61.
3. Briain RW. Back-translation for cross-cultural research. J Cross Cult Psychol 1970;1:185-216.
4. Singh B, Pandey N, Mehrotra B, Srivastava A, Chowdhury A, Tiwari S. Development of a comprehensive satisfaction index (ComSI) and its association with WHOQOL-BREF. Indian J Psychol Med 2019;6;562-8.
5. Hinkin TR. A Brief tutorial on the development of measures for use in survey questionnaires. Org Res Methods 1998;1:104-21.
6. Jollife IT. Principal Component Analysis. New York, NY: Springer; 1986.
7. Churchill GA. A Paradigm for developing better measures of marketing constructs. J Marketing Res 1979;16:64.
8. Braeken J, van Assen MA. An empirical Kaiser criterion. Psychol Methods 2017;22:450-66.
9. Cattell RB. The Scree test for the number of factors. Multivariate Behav Res 1966;1:245-76.
Menon and Praharaj: Development of rating scales

10. Bedi J, Verma T. Development of a scale of positive temperament in the Indian context. Indian J Psychol Med 2019;6:569-77.
11. Goodman R. The strengths and difficulties questionnaire: A research note. J Child Psychol Psychiatry 1997;38:581-6.

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.

How to cite this article: Menon V, Praharaj SK. Translation or development of a rating scale: Plenty of science, a bit of art. Indian J Psychol Med 2019;41:503-6.