Measurement validity in cross-cultural comparative research

MARTIN PRINCE

Institute of Psychiatry, King's College London

SUMMARY. **Background** – The purpose of this article is to review the procedures to establish measurement validity in cross-cultural comparative research, including recent developments in the quantitative assessment of cross-cultural construct validity. **Methods** – A narrative review, illustrated by selected examples, of methods in four areas – formative conceptual research, translation and adaptation, criterion validity and construct validity. **Results** – Valid assessment across cultures requires qualitative research to investigate the cultural relevance of the construct, a careful translation and adaptation of a common measure, followed by pre-testing and cognitive interviews on the populations to be tested. Full criterion validation across diverse cultures may be a chimera given the difficulty in establishing a universally applicable 'gold standard'. Quantitative analyses can, however, have a part to play in establishing construct validity across cultures. Scale internal consistency, inter-item and item-total correlations and test-retest reliability provide basic support for the viability of a measure in a new cultural setting. Exploratory factor analysis can be used to compare factors and factor loadings. The hypothesis of 'measurement invariance' across countries and cultures can be tested explicitly using confirmatory factor analysis (common underlying factors and factor loadings) and Rasch models (common hierarchality of items). Despite measurement invariance, threshold effects arising from cultural differences in norms, or expectations, or expressions of mental distress may still be a problem. **Conclusions** – There are few examples in the cross-cultural mental health literature of demonstrably valid culture-fair comparison. Much more, could, in principle, be done either to demonstrate measurement invariance, or to identify and explore sources of heterogeneity.

Declaration of Interest: None.

KEY WORDS: cross cultural comparison, psychometrics, measurement, validity.

Received 14.04.2008 – Final version received 17.04.2008 – Accepted on 19.04.2008.

Much 'cross-cultural' psychiatric research involves nothing more than the adaptation and application of concepts and methods developed in one cultural setting (usually the UK or the US) to another, typically a non-English speaking low or middle income country (LAMIC) in Latin America, Africa, or Asia. It is perhaps important to acknowledge at the outset that the existence of such research is, in the main, a reflection of the dominance of western biomedical and social science in terms of academic institutions, human resource, research funding and access to outlets for dissemination. English researchers, after all, do not conduct 'cross-cultural' research in the USA. This is not to critique its valuable contribution both to capacity building, and to our understanding of the burden of mental disorder in LAMIC. True cross-cultural comparative research is less common, much of it carried out under the aegis of the World Health Organisation (WHO); the *International Pilot Study of Schizophrenia* (Sartorius *et al.*, 1977; 1972); the *Collaborative Study on Strategies for Extending Mental Health Care* (Harding *et al.*, 1983); the *Collaborative Project on Psychological Problems in General Health Care* (PPGHC) (Sartorius *et al.*, 1993). More recent examples include the World Mental Health Survey of over 200,000 participants in two dozen countries (Demyttenaere *et al.*,2004; Kessler *et al.*, 2006), and the 10/66 Dementia Research Group's studies of dementia and ageing in ten LAMIC (Prince *et al.*, 2007). The purpose of this article is to review the necessary procedures to establish measurement validity across cultures, including some recent methodological developments in the quantitative assessment of cross-cultural construct validity.

In all such research, the need to establish the relevance of the constructs studied to the diverse settings, to use bicultural and bilingual translators to arrive at an appropriate local language version of the assessment, and to make all necessary cultural adaptations is well established; the methods and procedures have been extensively described, and a fairly clear consensus has emerged

Address for correspondence: Professor M. Prince, Institute of Psychiatry, King's College London P060, De Crespigny Park, London SE5 8AF (United Kingdom).
Fax: +44-20-78480137
E-mail: m.prince@iop.kcl.ac.uk

https://doi.org/10.1017/S1121189X00001305 Published online by Cambridge University Press

(Patel, 2003). Early descriptions of the process defined parameters by which a translated/ adapted questionnaire should be evaluated (Flaherty *et al.*, 1988):

1. Content equivalence: Is the construct meaningful, acknowledged, recognised in the culture?
2. Semantic equivalence: Does the meaning of each item remain the same after translation?
3. Technical equivalence: Does the method of data collection affect results differentially between two cultures? (This is highly likely to arise when adapting an assessment intended for literate educated western participants to participants in LAMIC. Self-completion questionnaires are not feasible; they could be interviewer administered but this could change their measurement properties. Complex response formats; for example, not at all/ several days/ more than half of the days/ nearly every day; may need to be simplified or broken down into components for verbal administration)
4. Criterion equivalence: What is an instrument's relationship with previously established and independent criteria for the same phenomena?
5. Conceptual equivalence: Does the instrument measure the same construct in different cultures?

This framework has been instrumental in raising awareness of cultural influences on idioms of distress, explanatory models, help-seeking behaviour and clinical responses – all of which have clear implications for nosology and measurement. However, the five sub-categories subsume three relatively distinct elements in terms of the research that is necessary to demonstrate measurement equivalence – the investigation, through formative research of the cultural relevance of the concept (content equivalence), the process of adaptation and translation of an existing measure (semantic and technical equivalence), and the local and cross-cultural validation of the measure at the end of the instrument adaptation/ development process (criterion and conceptual equivalence). These are now considered, in turn.

FORMATIVE RESEARCH TO INVESTIGATE THE CONSTRUCT UNDER STUDY

Kleinman first drew attention to the danger of committing a 'category fallacy', which he defined as "the reification of a nosological category developed for a particular cultural group that is applied to members of another culture for whom it lacks coherence" (Kleinman, 1987). To use his evocative example, the syndrome of 'soul loss' identified in some non-western settings could be operationalised and applied to an urban middle class north American population, and prevalence data would be generated, although clearly without validity. Kleinman's contention was that such fallacies were common but, given the dominance of the western cultural perspective, proceeded always in the opposite cultural direction. Errors might arise through the confusion of culturally distinctive behaviour with psychopathology on the basis of superficial phenomenological similarities. Hence, dysthymic disorder, he argued, might be inappropriately diagnosed in a low income country setting among those with chronic physical disorders – for example anaemia, nutritional or parasitic diseases – that were common in those settings but rare in north America where the DSM classification had been devised.

The coherence of western diagnostic paradigms was investigated in the seminal WHO Collaborative Study on Strategies for Extending Mental Health Care in Sudan, northern India and the Philippines (Wig *et al.*, 1980). Skeleton vignettes representing epilepsy, mental retardation, acute psychosis, mania, depressive psychosis, "process" schizophrenia and depressive neurosis were constructed and embellished by local teams using local idiom, events and customs. These were presented to multiple key informants. Each condition was widely recognised, epilepsy, mental retardation, psychosis and depressive neurosis being most commonly cited. All tended to be rated as serious, while the psychoses were thought particularly likely to lead to social and occupational disability. This study did not explore illness perceptions – similar vignette based (Patel & Prince, 2001) and qualitative studies (Shaji *et al.*, 2002; Cohen, 1995) in India indicated that while something resembling the dementia syndrome was widely recognised and indeed named, it was viewed as a normal part of ageing rather than as a medical condition that might lead to help-seeking. More 'bottom-up' ethnographic explorations of illness models include Manson's extension of the western Diagnostic Interview Schedule to include an operationalised list of symptoms and categories from Hopi indigenous nosology with a prima facie resemblance to depression (Manson *et al.*,1985), and the work of Patel and colleagues on Shona idioms for psychological distress in Zimbabwe (Patel *et al.*, 1995; 1997).

PROCESS OF TRANSLATION AND ADAPTATION OF INSTRUMENTS

The World Health Organisation has developed a protocol for translations of its English language assessments

https://doi.org/10.1017/S1121189X00001305 Published online by Cambridge University Press

to be approved for use in other settings (World Health Organisation, 2008). This probably represents 'state of the art' at present. The overall aim is defined succinctly as "to achieve different language versions of the English instrument that are conceptually equivalent in each of the target countries/cultures. That is, the instrument should be equally natural and acceptable and should practically perform in the same way. The focus is on cross-cultural and conceptual, rather than on linguistic/literal equivalence".

1. *Forward translation*

The translator should be bilingual. Their mother tongue should be that of the target culture, but they should also have knowledge of the culture from which the assessment originated. They should have technical expertise of the area covered by the instrument. Translators should consider the meaning of the original term and attempt to translate it aiming for conceptual equivalence rather than a word-for-word translation. They should strive to be simple, clear and concise. Fewer words are better. Long sentences with many clauses should be avoided. They should consider the typical respondent for the instrument being translated and what the respondent will understand when they hear the question, avoiding jargon and technical terms that cannot be understood clearly. They should consider gender and age applicability and avoid any terms that might be considered offensive.

2. *Expert panel*

A bilingual expert panel should be convened to identify and resolve inadequate expressions, as well as any discrepancies between the forward translation and the original. The expert panel may question words or expressions and suggest alternatives. In general, the panel should include the original translator, experts in health, as well as experts with experience in instrument development and translation. The result of this process will be a complete translated version of the questionnaire.

3. *Back-translation*

The instrument is translated back to English by an independent translator, whose mother tongue is English but who has no knowledge of the original questionnaire. Emphasis is again on conceptual rather than linguistic equivalence. Discrepancies require further work – forward translations, discussion by the bilingual expert panel, and back translation, iterated as necessary until a satisfactory version is reached.

4. *Pre-testing and cognitive interviewing*

Pre-test respondents should include individuals representative of those who will be administered the questionnaire, with purposive recruitment to ensure that both genders, all relevant age groups and different socioeconomic groups are included. An experienced interviewer administers the instrument and the respondent is then systematically debriefed. In the debriefing (sometimes referred to as cognitive interviewing) respondents are asked what they thought each question was asking, whether they could paraphrase the question in their own words, and their rationale for the answer that they gave. Any difficulties with comprehension are noted, together with requests for clarification. Any words or expressions that were found to be unacceptable or offensive are noted.

5. *Final version*

Each stage should be appropriately documented, including the initial forward version, the recommendations by the expert panel, the back-translation, a summary of problems found during the pre-testing and the modifications proposed; and the final version.

In reality, few translations aspire to, let alone achieve this level of sophistication. Mumford's careful translation of the Hospital Anxiety and Depression Questionnaire into Urdu is a good example (Mumford *et al.*, 1991b). Other approaches have been advocated. Sumathipala and Murray (Sumathipala & Murray, 2001) used a panel of nine to independently translate items from the Bradford Somatic Inventory into Sinhala, and then to rate these for conceptual and semantic equivalence. Translations failing to reach consensus were discussed by the group for modifications and subjected to further discussion to achieve consensus. This combined the processes of forward translation and expert panel review. The authors argued that no single forward translation could hope to address satisfactorily the complexities of rendering problematic items into the new language, and that the consensus element did away with the need for back translation to ensure content and semantic equivalence. The utility of back translation has been questioned since misjudgement on the part of forward or back translators can result in a lack of sensitivity and specificity for the identification of problematic items. Moreover, while back translation may help to identify problems it cannot generate solutions.

CRITERION VALIDITY

Testing of criterion validity requires a 'gold standard', technically the very thing that one is setting out to measure. The most commonly used paradigm in psychiatric research is the clinician semi-structured interview, generally the *Schedules for Clinical Assessment in Neuropsychiatry* (SCAN) or the *Structured Clinical Interview for DSM Disorders* (SCID), applying ICD10 or DSM IV diagnoses. Kessler in his description of the validation procedures planned for the World Mental Health

https://doi.org/10.1017/S1121189X00001305 Published online by Cambridge University Press

Survey draws our attention to several problems that are implicit in this process, most of which are not particular to cross-cultural research (Kessler *et al.*, 2004). First, the reliability of clinician semi-structured interviews is by no means perfect, particularly in community-based research, and this random error will set an upper limit on the validity coefficients that are likely to be observed. Second, repetition of comprehensive mental state assessments has been shown to be associated with systematic underreporting of symptoms on the second interview compared with the first, again tending towards an underestimate of true validity. Third, the DSM and ICD diagnostic criteria themselves are not fully operationalised, and differing judgments made in the algorithms accompanying the test assessment and gold standard research interviews may be another source of discrepancy. Fourth, Cohen's kappa varies with the prevalence of the disorder even when specificity and sensitivity are constant, limiting the utility of this validity coefficient in comparing the validity of a measure across different populations – Kessler proposes instead the area under the ROC curve, which in the special case of a dichotomous test variable is the average of sensitivity and specificity (Kessler *et al.*, 2004).

For all of these reasons Kessler suggested that it would be more appropriate instead to 'calibrate' the lay structured WMH-CIDI against the clinician semi-structured SCID (Kessler *et al.*, 2004). Prediction equations were developed in a clinical calibration subsample, with WMH-CIDI symptom-level data used to predict SCID diagnoses, and these coefficients were then used to assign predicted probabilities of SCID diagnoses to each respondent in the remainder of the sample. It was then possible to investigate whether estimates of prevalence and associations based on WMH-CIDI diagnoses were consistent with those based on predicted SCID diagnoses. With data published so far for WMH surveys in France, Italy, Spain, and the US, this was the case for 12-month prevalence estimates of anxiety and mood disorder, while lifetime prevalence estimates from the CIDI were conservative with respect to the SCID (Haro *et al.*, 2006). Diagnostic concordance was also good for any anxiety disorder (AUC = 0.88) and any mood disorder (AUC = 0.83). No information is yet available on the validity or calibration of the WMH-CIDI diagnoses in WMH centres in LAMIC in Asia, Africa and Latin America.

The 10/66 Dementia Research Group used a similar calibration approach to arrive at its '10/66 Dementia' diagnosis. However, the calibration algorithm was first developed and tested in an international pilot study conducted in 26 centres in Latin America, Africa, India and China (Prince *et al.*, 2003). The assessments that contributed to the algorithm were selected empirically on the basis of culture-fair characteristics (capable of discriminating equivalently between DSM IV dementia cases and non-cases across a variety of settings) and education-fair characteristics (capable of discrimination across the full range of educational abilities, and with a minimal false positive rate among those with little or no education). The predictive algorithm was developed on one random half of the pilot data set, and tested on the other. The resulting probabilistic 10/66 diagnosis showed similar validity in all world regions, with 94% sensitivity overall, and false positive rates of 6% and 3% in low and high education controls, respectively (Prince *et al.*, 2003). The algorithm was then ready to be applied in the subsequent population-based studies, at which stage it would be subject to further concurrent and predictive validation (Prince *et al.*, 2007).

A more fundamental concern with respect to cross-cultural research, is whether it is ever really possible to identify a single 'gold standard' that could be applied across a wide variety of cultural settings. This is not properly addressed in either the WMH studies or 10/66 studies. The SCID, for example has not itself been validated against locally appropriate criteria, and the universal applicability of the DSM IV nosologies has also been challenged (Mezzich *et al.*, 1999). The demonstration of equivalent validity against a universal external criterion does not exclude the commission of a category fallacy as described by Kleinman (1987). There are examples in the literature of the use of external criteria that may more accurately reflect local diagnostic practice. For example, clinician checklists can be used that are consistent with standard nosologies, but at least allow local clinicians to use culturally appropriate interview techniques to elicit and rate symptoms that they consider to be relevant. This approach was used successfully in Sri Lanka to validate the *Revised Clinical Interview Schedule* (CIS-R) at the symptom cluster level (Wickramasinghe *et al.*, 2002). In Ethiopia, the *Comprehensive Psychopathological Rating* Scale (CPRS) was used as the gold standard to validate the *Self-Reporting Questionnaire* and the *Edinburgh Postnatal Depression Scale*, after its inter-rater and test-retest reliability had first had been assessed and improved through training (Hanlon *et al.*, 2007). In an innovative approach, Patel used recognition by local practitioners (both allopathic and traditional healers), when confirmed by a western structured psychiatric interview (the CIS-R), as the gold standard against which to validate the Shona Symptom Questionnaire as an assessment of common mental disorders (Patel *et al.*, 1997). There are no examples yet of such approaches being used in validation studies across a variety of countries and cultures. One

https://doi.org/10.1017/S1121189X00001305 Published online by Cambridge University Press

possible avenue to explore in the future would be to use both a common 'etic' criterion (such as the SCAN) together with 'emic' criteria that could be selected for each cultural setting. The reliability of the criterion is clearly a key consideration.

CONSTRUCT VALIDITY

Construct validity is generally not well understood, but is highly relevant to establishing the validity of constructs and assessments across populations and cultures. The term was proposed by the APA Committee on Psychological Tests in 1954, including Meehl and Cronbach who wrote a classic paper defining the term (Cronbach & Meehl, 1955). Pre-existent concepts of validity had focussed on content validity and criterion-related validity, where a gold-standard for the construct was already established. Construct validity, as defined by Meehl and Cronbach, referred to a situation in which 'the tester has no definite criterion measure of the quality with which he is concerned, and must use indirect measures. Here the (elucidation of the) trait or quality underlying the test if of central importance'. Meehl and Cronbach considered that expert ratings of these matters would be of limited if any value. Construct validity would be involved in answering such questions as: To what extent is this test culture-free? Does this test measure reading ability, quantitative reasoning, or response sets? How does a person with a high score differ from a person with a low score? The answers to these questions would be derived from quantitative research, essentially through a series of hypothesis-driven investigations aimed at identifying the 'nomological net' or theoretical framework consisting of more or less proximate identifiers for the construct, at least some of which would need to be observable.

Meehl and Cronbach described the investigation of a test's construct validity as comparable to that for developing and confirming theories in observational research. This was particularly the case for test- "criterion" correlations used to define the nomological net (in inverted commas since the construct validity of the criteria themselves is also under investigation, and, as with a hypothesis in observational research, the construct that is intended to be measured can be refined on the basis of findings). Other quantitative procedures for establishing construct validity included inter-item correlations and internal consistency, factor analysis and test-retest reliability, Methodological developments in recent years have led to the introduction of more sophisticated, hypothesis-driven approaches to testing construct validity across cultures. Confirmatory Factor Analysis (CFA) and Rasch models can be used to assess whether a scale measures the same trait dimension, in the same way, when applied in qualitatively distinct groups; this is referred to as 'measurement invariance'. In a worked example, Reise *et al.* (1993) illustrated these approaches with respect to mood symptoms assessed in Minnesota and China). The demonstration of measurement invariance across countries, cultures and/ or ethnic groups provides strong evidence to support the construct validity of an assessment for the purposes of comparative research. Despite their evident applicability to cross-cultural validity of assessments of mental health status, there are relatively few examples in the literature.

Confirmatory Factor Analysis (CFA)

Principal Component Analysis (PCA) is an exploratory technique based on a covariance matrix of tetrachoric correlations to estimate factor loadings of items upon underlying latent traits (Olsson, 1979). PCA has been widely used to demonstrate similar patterns of factor loadings when a measure is used in different cultural contexts – for example the Bradford Somatic Inventory used on patients with medically unexplained symptoms in general healthcare settings in Pakistan and the UK (Mumford *et al.*, 1991a). In Confirmatory Factor Analysis a model is tested that specifies in advance the relations between observed variables and latent factors. Testing for measurement invariance (Sorbom, 1974) involves an assessment of whether the best factor solution relates to the latent trait or traits in the same way in each of several populations. CFA has been used to validate the Parental Bonding Interview across European centres participating in the World Mental Health Survey (Heider *et al.*, 2005), providing strong evidence for the invariance of a three factor solution suggested by exploratory factor analysis, across six countries, both genders and all age groups. The general approach is to compare two models in which all item loadings are i) constrained and ii) not constrained to be identical between countries. Partial measurement invariance may be tested for if full measurement is not – supported. Partial measurement invariance occurs if a majority of the non-fixed values are still invariant across groups and if these invariant loadings define the latent metric. The absolute fit of a CFA model can be evaluated by means of a ÷ 2 statistic (Bentler & Bonett, 1980). Several other absolute and relative indices of fit are recommended. Akaike's Information Criterion (AIC) (Akaike, 1987) adjusts the model chi-square to penalize

https://doi.org/10.1017/S1121189X00001305 Published online by Cambridge University Press

for model complexity. The lower the AIC value, the better the fit of the model (Burnham & Anderson, 1998). The Tucker-Lewis index (TLI)(Tucker & Lewis, 1973) indicates the proportion of covariation among indicators explained by the model relative to a null model of independence, and is independent of sample size. Values near zero indicate poor fit, whereas values near 1.0 indicate good fit; those greater than 0.90 are considered satisfactory (Marsh *et al.*, 1993; Dunn *et al.*, 1993). In contrast, the Root Mean Square error of Approximation (RMSEA) assesses badness of fit per degree of freedom in the model and zero if the model fits perfectly; RMSEA values of less than 0.05 indicate close fit and 0.05 to 0.08 reasonable fit of a model (Browne, 1990).

Rasch Models

Rasch analysis belongs to a family of statistical models developed from item response theory (IRT) (Reise *et al.*, 1993; Nunnally & Bernstein, 1994). Whereas CFA models account for the covariance between test items, IRT models account for patterns of item responses. The Rasch model (Rasch, 1993) suggests that the responses to a set of items can be justified by a person's position on the underlying trait that is being measured and by the item severity (also referred to as item calibration). It is measured on the same scale as the severity of the underlying trait. Participants with severity scores below the calibration for the given item are more likely to deny than endorse the item, while those with severity scores above the calibration for the item are more likely to endorse it. It follows that the items with higher calibrations are less frequently endorsed. Also, a participant who endorses an item of middling severity is likely to have endorsed all items with lower calibrations. INFIT assesses the extent to which observed response patterns were consistent, or inconsistent with the item calibrations. Mean square INFIT (MNSQ) statistics of < 1.3 indicate that the subscale items contribute to a single hierarchical underlying construct (unidimensionality). While there can be no single satisfactory test of model adequacy, techniques can be used to assess different aspects of goodness of fit of Rasch models.

Rasch analysis was used in the WHO international study of Psychological Problems in General Health Care settings (PPGHC) (Simon *et al.*, 2002). Strong evidence was provided to support a common hierarchality for CIDI symptoms (in order of item severity; sleep, mood, concentration fatigue, loss of interest, suicidality, guilt, agitation/ retardation, appetite change) across three groups of countries categorized by level of prevalence of major depression (low/ intermediate/ high). This impression was confirmed by a formal test for differential item functioning; the goodness of-fit for a model allowing different symptom thresholds for the three prevalence groups was not significantly better than that for a simpler model assuming identical symptom thresholds (Chi sq 19.4, df 18, P=0.38). The authors concluded from these findings that there was no evidence that prevalence differences were explained by cross-cultural differences in the form or validity of the depressive syndrome, and that therefore the apparent differences in depression prevalence could not be attributed to 'category fallacy'.

AN EXAMPLE OF CROSS-CULTURAL CONSTRUCT VALIDATION – THE EURO-D

Background

The EURO-D scale (Prince *et al.*, 1999b) was originally developed in an effort to harmonize data on late-life depression from population-based studies in 11 European countries as part of the EURODEP collaboration. EURO-D items were all taken from the Geriatric Mental State (Copeland *et al.*, 1986). However, they were also used in several of the other assessments to be harmonized – the SHORT-CARE, the Centre for Epidemiological Studies Depression scale (CES-D) the *Zung Self-Rating Depression Scale and Comprehensive Psychopathological Rating Scale* (CPRS) (Prince *et al.*, 1999b). Thus, the EURO-D has intrinsically strong face validity. The initial validation showed adequate internal consistency. The optimal cut-off point was 3/4 for the external criteria of both GMS depression and SHORT-CARE pervasive depression (Prince *et al.*,1999b). The EURO-D was later found to be reliable and valid for detection of DSM-IIIR depression in older people in Spain (Larraga *et. al.*, 2006). In the EURODEP studies, principal components analysis generated two factors (Affective Suffering and Motivation) that were common to nearly every participating country (Prince *et al.*, 1999b). This factor structure was later confirmed for Indian, Latin-American and Caribbean centers in the 10/66 Dementia Research Group pilot studies (Prince *et. al.*, 2004).

Cross-cultural measurement invariance

A more detailed investigation of the cross-cultural validity of the EURO-D and its measurement properties

https://doi.org/10.1017/S1121189X00001305 Published online by Cambridge University Press

(Castro-Costa *et al.*, 2008) was carried out using data from the Survey of Health, Ageing and Retirement in Europe (SHARE) 2004 baseline study (Borsch-Supan *et al.*, 2005). This comprised cross-sectional surveys of representative samples of community residents aged 50 years and over from 10 European countries (Denmark, Sweden, Austria, France, Germany, Switzerland, the Netherlands, Spain, Italy and Greece). Principal Component Analysis again suggested a two-factor solution (Affective Suffering and Motivation) in 9 of the 10 countries. In all countries, the full EURO-D scale was moderately internally consistent with Cronbach's alpha ranging from 0.62 to 0.78. For the confirmatory factor analysis (Castro-Costa *et al.*, 2008) the fit of three two-factor solutions (Affective Suffering and Motivation) was tested – Model 1: with no constraints; Model 2: both factors were constrained to load equally across all countries, and Model 3: only items from Affective Suffering factor were constrained. According to all criteria, each of the three two-factor solution models fitted better than a one-factor solution. Comparison between Model 2, with all items loadings constrained, and Model 1 showed that the constraints increased the model chi square by 737.4 with 81 additional degrees of freedom. More importantly, the AIC value for Model 2 (2353.3), was higher than the AIC model for Model 1 (1777.9), and the goodness-of-fit indices (TLI: 0.88, RMSEA: 0.016) were slightly worse than those for Model 1 (TLI: 0.89, RMSEA: 0.015). Therefore, the full invariance model did not hold across countries, and was rejected. In the partial invariance Model (Model 3), specified after closer inspection of Model 1, the factor loading parameters of Motivation items were freely estimated, while those for the Affective Suffering items were constrained. While the chi-square change (412.5 on 54 degrees of freedom) still suggests that Model 1 is superior, the three absolute goodness of fit indices provided strong evidence that the partial invariance Model (model 3 – AIC: 2082.4, TLI: 0.89, RMSEA: 0.015) was adequate and did not clearly differ from Model 1, which was without any restriction.

The Rasch model indicated that the EURO-D was a hierarchical scale (Castro-Costa *et al.*, 2008). The rank ordering of EURO-D item calibration values was similar for all countries. Depression (followed by sleep disturbance) had the lowest item calibrations overall, while guilt and suicidality had the highest. Spearman correlation coefficients between the set of EURO-D item calibrations for pairs of SHARE project countries ranged from 0.70 to 0.97. The intraclass correlation coefficient for agreement in item calibrations across all 10 countries was 0.89 (0.78-0.95). There was therefore strong evidence for measurement invariance with respect to item calibration. However, there was more heterogeneity in item calibration between countries for the Motivation than for the Affective Suffering items. For the subset of items loading on Affective Suffering the ICC was 0.94 (0.85-0.99) whereas for the four Motivation items it was only 0.65 (0.30-0.96).

What are the factors measuring? The nomological net

It had been previously noted in both the EURODEP (Prince *et al.*, 1999a) and the 10/66 pilot studies (Prince *et al.*, 2004) that the Affective Suffering factor was associated with female gender, while the Motivation factor scores, but not Affective Suffering scores, increased with increasing age. Did this correlation between Motivation and age, considered in the context of the orthogonal relationship with Affective Suffering suggest that an expressed lack of interest and enjoyment, together with pessimism might be affectively neutral statements representing an adaptive cognitive appraisal of activity limitation in older age? An alternative hypothesis was that the Motivation factor might be capturing some of the clinical features of vascular depression (Alexopoulos *et al.*, 1997) – impaired executive functioning associated with vascular damage to sub-cortical structures. In the SHARE data set the previously observed pattern of association with age and gender, was confirmed consistent across all 10 SHARE countries (Castro-Costa *et al.*, 2007). We were now able to test the earlier hypothesis regarding impaired executive functioning and did indeed find that higher Motivation factor scores were associated with poorer performance on animal naming (executive function) while there was no association with 10 word recall (memory). The Affective Suffering was factor was not associated with either cognitive test. Again, the pattern of association between the cognitive tests and the EURO-D factors was highly consistent across all countries (Castro-Costa *et al.*, 2007).

Summary of findings

1. In all settings, the EURO-D held up well as a unidimensional scale as evidenced by the moderately high Cronbach's alpha and the reasonable fit of the one factor solution.
2. However, there was also evidence that it may be measuring two underlying factors, Affective Suffering (well characterized and invariant across cultures) and

https://doi.org/10.1017/S1121189X00001305 Published online by Cambridge University Press

Motivation (less well characterized and variable across cultures).
3. The Affective Suffering factor showed excellent cross-cultural measurement properties with strong evidence for measurement invariance with respect both to item loadings onto a common underlying latent trait, and to item calibrations. It has strong face validity as a depression measure.
4. The Motivation factor exhibited considerable heterogeneity in factor loading patterns and item calibrations between countries, and was not sufficiently well characterized or stable for use in cross-cultural research.
5. The two factors showed a different but well-characterised pattern of association with age, gender and cognitive impairment, that was highly cross-culturally consistent. The associations between Motivation and older age and impaired executive function raised legitimate concerns about its relevance to the core depressive syndrome. The construct needs to be further clarified, and its clinical and predictive validity established. Other testable hypotheses are indicated in Table I.

Table I – *Establishing the 'nomological net' for the EURO-D scale Affective Suffering and Motivation factors.*

	Affective Suffering	Motivation
Previously tested associationk		
Age	0	+
Education	0	0
Female gender	+	0
Memory	0	0
Executive function	0	–
Further hypotheses to be tested		
Periventricular white matter lesion	0	+
Disability	0/+	++
Mortality	+	++
Deliberate self harm and suicide	+	0
Past history of depression	+	0
Life events	+	0
Response to antidepressants	+	0

+ Positive association
– Negative association
0 No association

CONCLUSION

In summary, a basic checklist of steps to be taken to achieve valid assessment across cultures should include:

1. qualitative research to investigate the cultural relevance of the construct and the local idiom
2. a careful translation and adaptation
3. pre-testing and cognitive interviews on the populations to be tested.

Full criterion validation across diverse cultures may be a chimera. Quantitative analyses can, however, have an important part to play in establishing the construct validity of measures across cultures, ethnicities and populations. Scale internal consistency and inter-item and item-total correlations and test-retest reliability provide basic support for the viability of a measure in a new cultural setting. Exploratory factor analysis can be used to compare factors and factor loadings with those from other cultural settings where there is more experience in the use of the measure. The hypothesis that there is 'measurement invariance' across countries and cultures can be tested explicitly using confirmatory factor analysis (common underlying factors and factor loadings) and Rasch models (common hierarchality of items). Despite measurement invariance, threshold effects arising from culturally determined differences in norms or expectations or expressions of mood and mental health may still be a problem (Simon *et al.*, 2002), and may explain much of the observed difference in prevalence. Such influences have been clearly established in the case of self-reported global health, where the use of anchoring vignettes has been advocated to identify and adjust for the consequent response bias (Salomon *et al.*, 2004). In principle, a similar approach might improve the cross-cultural comparability of self-reported mental health assessments.

REFERENCES

Akaike H. (1987). Factor analysis and AIC. *Psychometrika* 52, 317-332.

Alexopoulos G.S., Meyers B.S. & Young R.C. (19979. Vascular depression hypothesis. *Archives General Psychiatry* 54, 915-922.

Bentler P.M. & Bonett D.G. (1980). Significance tests and goodness of fit in the analysis of covariances structures. *Psychological Bulletin* 88, 588-606.

Borsch-Supan A., Brugiavini A., Jurges H., Mackenbach J., Siegrist J. & Weber G. (2005). *Health, Ageing and Retirement in Europe – First Results from the Survey of Health, Ageing and Retirement in Europe*. MEA: Manheim.

Browne M.W. (1990). *MUTMUM PC: User's Guide* Ohio State University: Columbus.

Burnham K.P. & Anderson D.R. (1998). *Model Selection and Inference: A Practical Information-Theoretic Approach*. Springer-Verlag: New York.

Castro-Costa E., Dewey M., Stewart R., Banerjee S., Huppert F., Mendonca-Lima C., Bula C., Reisches F., Wancata J., Ritchie K., Tsolaki M., Mateos R. & Prince M. (2007). Prevalence of depressive symptoms and syndromes in later life in ten European countries: the SHARE study. *British Journal of Psychiatry* 191, 393-401.

Castro-Costa E., Dewey M., Stewart R., Banerjee S., Huppert F., Mendonca-Lima C., Bula C., Reisches F., Wancata J., Ritchie K., Tsolaki M., Mateos R. & Prince M. (2008). Ascertaining late-life depressive symptoms in Europe: an evaluation of the survey version of the EURO-D scale in 10 nations. The SHARE project. *International Journal of Methods in Psychiatric Research* 17(1), 12-29.

https://doi.org/10.1017/S1121189X00001305 Published online by Cambridge University Press

Cohen L. (1995). Toward an anthropology of senility: anger, weakness, and Alzheimer's in Banaras, India. *Medical Anthropology Quarterly* 9(3), 314-334.

Copeland J.R., Dewey M.E. & Griffiths-Jones H.M. (1986).Computerised psychiatric diagnostic system and case nomenclature for elderly subjects:GMS and AGECAT. *Psychological Medicine* 16, 89-99.

Cronbach L.E. & Meehl P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin* 52, 281-302.

Demyttenaere K., Bruffaerts R., Posada-Villa J., Gasquet I., Kovess V., Lepine J. P., Angermeyer M.C., Bernert S., de Girolamo G., Morosini P., Polidori G., Kikkawa T., Kawakami N., Ono Y., Takeshima T., Uda H., Karam E.G., Fayyad J.A., Karam A.N., Mneimneh Z.N., Medina-Mora M.E., Borges G., Lara C., de Graaf R., Ormel J., Gureje O., Shen Y., Huang Y., Zhang M., Alonso J., Haro J.M., Vilagut G., Bromet E.J., Gluzman S., Webb C., Kessler R.C., Merikangas K.R., Anthony J.C., von Korff M.R., Wang P.S., Brugha T.S., Guilar-Gaxiola S., Lee S., Heeringa S., Pennell B.E., Zaslavsky A.M., Ustun T.B. & Chatterji S. (2004). Prevalence, severity, and unmet need for treatment of mental disorders in the World Health Organization World Mental Health Surveys. *Journal of American Medical Association* 291(21), 2581-2590.

Dunn G., Everitt B. & Pickles A. (1993). *Modelling Covariances and Latent Variables using EQS*, 1a ed. Chapman & Hall: London.

Flaherty J.A., Gaviria F.M., Pathak D., Mitchell T., Wintrob R., Richman J.A. & Birz S. (1988). Developing instruments for cross-cultural psychiatric research. *Journal of Nervous and Mental Disease* 176(5), 257-263.

Hanlon C., Medhin G., Alem A., Araya M., Abdulahi A., Hughes M., Tesfaye M., Wondimagegn D., Patel V. & Prince M. (2007). Detecting perinatal common mental disorders in Ethiopia: Validation of the self-reporting questionnaire and Edinburgh Postnatal Depression Scale. *Journal of Affective Disorders* 108, 251-262.

Harding T.W., Climent C.E., Diop M., Giel R., Ibrahim H.H., Murthy R.S., Suleiman M.A., & Wig N.N. (1983). The WHO collaborative study on strategies for extending mental health care, II: The development of new research methods. *American Journal of Psychiatry* 140(11), 1474-1480.

Haro J.M., Arbabzadeh-Bouchez S., Brugha T.S., de Girolamo G., Guyer M.E., Jin R., Lepine J.P., Mazzi F., Reneses B., Vilagut G., Sampson N.A. & Kessler R.C. (2006). Concordance of the Composite International Diagnostic Interview Version 3.0 (CIDI 3.0) with standardized clinical assessments in the WHO World Mental Health surveys. *International Journal of Methods in Psychiatric Research* 15(4), 167-180.

Heider D., Matschinger H., Bernert S., Vilagut G., Martinez-Alonso M., Dietrich S. & Angermeyer M.C. (2005). Empirical evidence for an invariant three-factor structure of the Parental Bonding Instrument in six European countries. *Psychiatry Research* 135(3), 237-247.

Kessler R.C., Abelson J., Demler O., Escobar J.I., Gibbon M., Guyer M.E., Howes M.J., Jin R., Vega W.A., Walters E.E., Wang P., Zaslavsky A. & Zheng H. (2004). Clinical calibration of DSM-IV diagnoses in the World Mental Health (WMH) version of the World Health Organization (WHO) Composite International Diagnostic Interview (WMHCIDI). *International Journal of Methods in Psychiatric Research* 13(2), 122-139.

Kessler R.C., Haro J. M., Heeringa S.G., Pennell B.E. & Ustun T.B. (2006). The World Health Organization World Mental Health Survey Initiative. *Epidemiologia e Psichiatria Sociale* 15(3), 161-166.

Kleinman A. (1987). Anthropology and psychiatry. The role of culture in cross-cultural research on illness. *British Journal of Psychiatry* 151, 447-454.

Larraga L., Saz P., Dewey M.E., Marcos G. & Lobo A. (2006). Validation of the Spanish version of the EURO-D scale: an instrument for detecting depression in older people. *International Journal of Geriatric Psychiatry* 21(12), 1199-1205.

Manson S.M., Shore J.H. & Bloom J.D. (1985). The depressive experience in American Indian communities. A challenge for psychiatric theory and diagnosis. In *Culture and Depression* (ed. A. Kleinman and B. Good), pp. 331-368. University of California Press: Berkeley.

Marsh H.W., Balla J.R. & Hau K.T. (1996). An evaluation of incremental fit indices: a clarification of mathematical and empiracal properties. In A*dvanced Structural Equation Modelling: Issues and Techniques* (ed. G.A. Marcoulides and R.E. Schumacker), pp. 315-355. Lawrence Erlbaum Associates: Mahwah.

Mezzich J.E., Kirmayer L.J., Kleinman A., Fabrega H., Jr., Parron D.L., Good B.J., Lin K.M. & Manson S.M. (1999). The place of culture in DSM-IV. *Journal of Nervous and Mental Disease* 187(8), 457-464.

Mumford D.B., Bavington J.T., Bhatnagar K.S., Hussain Y., Mirza S. & Naraghi M.M. (1991a). The Bradford Somatic Inventory. A multi-ethnic inventory of somatic symptoms reported by anxious and depressed patients in Britain and the Indo-Pakistan subcontinent. *British Journal of Psychiatry* 158, 379-386.

Mumford D.B., Tareen I.A., Bajwa M.A., Bhatti M.R. & Karim R. (1991b). The translation and evaluation of an Urdu version of the Hospital Anxiety and Depression Scale. *Acta Psychiatrica Scandinaviac* 83(2), 81-85.

Nunnally J.C. & Bernstein I.H. (1994). *Psychometric Theory*, 3rd ed. McGraw-Hill: New York.

Olsson U. (1979. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika* 44, 443-460.

Patel V. (2003). Cultural issues in measurement and research. In *Practical Psychiatric Epidemiology* (ed. M. Prince et al.). Oxford University Press: Oxford.

Patel V. & Prince M. (2001). Ageing and mental health in a developing country: who cares? Qualitative studies from Goa, India. *Psychological Medicine* 31(1), 29-38.

Patel V., Musara T., Butau T., Maramba P. & Fuyane S. (1995). Concepts of mental illness and medical pluralism in Harare. *Psychological Medicine* 25, 485-493.

Patel V., Simunyu E., Gwanzura F., Lewis G. & Mann A. (1997). The Shona Symptom Questionnaire: the development of an indigenous measure of common mental disorders in Harare: *Acta Psychiatrica Scandinavica* 95(6), 469-475.

Prince M., Beekman A., Fuhrer R., Hooijer C., Kivela S., Lawlor B., Lobo A., Magnusson H., Meller I., Oyen Reischies F., Skoog I., Turrina C. & Copeland J.R.M. (1999a). Depression symptoms in late-life assessed using the EURO-D scale. Effect of age, gender and marital status in 14 European centres. *British Journal of Psychiatry* 174, 339-345.

Prince M., Reischies F., Beekman A.T.F., Fuhrer R., Jonker C., Kivela S.L., Lawlor B., Lobo A., Magnusson H., Fichter M.M., Van Oyen H., Roelands M., Skoog I., Turrina C. & Copeland J.R. (1999b). Development of the EURO-D scale – a European Union initiative to compare symptoms of depression in 14 European centres. *British Journal of Psychiatry* 174, 330-338.

Prince M., Acosta D., Chiu H., Scazufca M. & Varghese M. (2003). Dementia diagnosis in developing countries: a cross-cultural validation study. *Lancet* 361, 909-917.

Prince M., Acosta D., Chiu H., Copeland J., Dewey M., Scazufca M. & Varghese M. (2004). Effects of education and culture on the validity of the Geriatric Mental State and its AGECAT algorithm. *British Journal of Psychiatry* 185, 429-436.

Prince M., Ferri C.P., Acosta D., Albanese E., Arizaga R., Dewey M., Gavrilova S.I., Guerra M., Huang Y., Jacob K.S., Krishnamoorthy E.S., McKeigue P., Rodrigues J.L., Salas A., Sosa A.L., Sousa R., Stewart R. & Uwakwe R. (2007). The protocols for the 10/66 Dementia Research Group population-based research programme. *BMC Public Health* 7, 165.

Rasch G. (1993). *Probabilistic Models for Some Intelligence and Attainment Tests*. Mesa Press: Chicago, IL.

Reise S.P., Widaman K.F. & Pugh R.H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin* 114(3), 552-566.

https://doi.org/10.1017/S1121189X00001305 Published online by Cambridge University Press

Salomon J.A., Tandon A., Murray C.J.L. & World Health Survey Pilot Study Collaborating Group (2004). Comparability of self rated health: cross sectional multi-country survey using anchoring vignettes. *British Medical Journal* 328(7434), 258.

Sartorius N., Shapiro R., Kimura M. & Barrett K. (1972). WHO international pilot study of schizophrenia. *Psychological Medicine* 2(4), 422-425.

Sartorius N., Jablensky A. & Shapiro R. (1977). Two-year follow-up of patients included in the WHO International Pilot Study of Schizophrenia. *Psychological Medicine* 7, 529-541.

Sartorius N., Ustun T.B., Costa E.S.J., Goldberg D., Lecrubier Y., Ormel J., Von Korff M. & Wittchen H.-U. (1993). An international study of psychological problems in primary care. Preliminary report from the World Health Organization Collaborative Project on Psychological Problems in General Health Care. *Archives of General Psychiatry* 50(10), 819-824.

Shaji K.S., Smitha K., Praveen Lal K. & Prince M. (2002). Caregivers of patients with Alzheimer's Disease : A qualitative study from the Indian 10/66 Dementia Research Network. *International Journal of Geriatric Psychiatry* 18, 1-6.

Simon G.E., Goldberg D.P., Von Korff M. & Ustun T.B. (2002).Understanding cross-national differences in depression prevalence. *Psychological Medicine* 32(4), 585-594.

Sorbom D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology* 27, 229-239.

Sumathipala A. & Murray J. (2001). New approach to translating instruments for cross-cultural research: a combined qualitative and quantitative approach for translation and consensus generation. *International Journal of Methods in Psychiatric Research* 9(2), 85-97.

Tucker L. & Lewis C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika* 38, 1-10.

Wickramasinghe S.C., Rajapakse L., Abeysinghe R. & Prince M. (2002). The Clinical Interview Schedule-Revised (CIS-R): modification and validation in Sri Lanka. *International Journal of Methods in Psychiatric Research* 11(4), 169-177.

Wig N.N., Suleiman M.A., Routledge R., Murthy R.S., Ladrido-Ignacio L., Ibrahim H.H. & Harding T.W. (1980). Community reactions to mental disorders. A key informant study in three developing countries. *Acta Psychiatrica Scandinavica* 61(2), 111-126.

World Health Organisation (2008). Process of translation and adaptation of instruments. Retrieved May 15, 2008, from http://www.who.int/substance_abuse/research_tools/translation/en/print.html

https://doi.org/10.1017/S1121189X00001305 Published online by Cambridge University Press