3mc2008 – Proceedings ©

A Systematic Approach to Test and Questionnaire Adaptations in an African Context

Penny Holding, PhD
Africa Mental Health Foundation, Kenya/ Case Western Reserve University, USA
penny.holding@uclmail.net

Amina Abubakar, PhD
Tilburg University, Netherlands/ Centre for Geographic Medicine Research (Coast), KEMRI, Kilifi/ Africa Mental Health Foundation, Kenya
A.AbubakarAli@uvt.nl

Patricia Kitsao-Wekulo MEd
Africa Mental Health Foundation, Kenya
wekulo@netathome.co.ke

A Systematic Approach to Test and Questionnaire Adaptations in an African Context

Sub-Saharan Africa, like most other resource-limited settings, lacks measures of childhood outcome that are appropriately standardized, validated and with documented reliability. This paper highlights issues that should be considered by researchers and clinicians when planning for psychological evaluations in contexts where measures have seldom, if ever, been developed. The strengths and weaknesses of different methodologies are discussed, and we outline a systematic approach to test adaptation formulated through more than a decade of experience of test development and application in Kenya (Holding, Taylor et al. 2004; Carter, Less et al. 2005; Abubakar, van de Vijver et al. 2008; Alcock, Holding et al. 2008; Holding and Kitsao-Wekulo under review)

The literature on the use of psychological tests in diverse cultures is rich in examples that illustrate the necessity of making modifications to test content and administration techniques (Serpell 1979; Kathuria and Serpell 1998; Holding, Taylor et al. 2004; Jukes, Pinder et al. 2006). The need to make modifications stems from fundamental cultural differences between test takers in North America and Europe, for whom the majority of assessments have been designed, and those from rural Africa. Without these modifications, the psychometric qualities of the test data may be questionable, and the distribution of scores elicited may show a lack of sensitivity to within-population differences. Based upon the premise that no instrument can claim to be culture-free (Scarr 1984; Nell 2000), the appropriateness of an instrument to a specific target population will need to be determined by evaluating its psychometric properties in context, as well as its cultural, developmental and health or educational relevance. Beyond the characteristics of the test taker, the attributes of the test administrator should also be considered. Few countries in Africa posses a psychological service, and as a consequence also lack training opportunities in test administration and interpretation. Research studies are therefore likely to be dependent upon technicians with limited experience to administer their test batteries.

Approaches to Test Development:

Three approaches to test development, Adoption, Adaptation, and Assembly, have been proposed to address the shortage of measures. The first approach, Adoption, involves taking in its entirety a test already in use in another population. The language used in the test may need to be translated into that of the new target population, but test content and procedures are administered as per the original standardisation. One advantage of this approach is that a tool is readily available for use. The second lies in the precedence afforded by an instrument that has been applied in comparative populations. It is frequently assumed that as long as it is only the language that has changed, the instrument remains standardised, and it will therefore be possible to directly compare the performance of the new target group to that of the standardisation sample. However, using standardisations that do not include members of the new target group in the original sample can lead to systematic selection bias, even when the test is being applied to minority groups residing within the same culture (Reynolds 1983). The direct adoption of measures from other cultural settings has been found to constrain within-population variance, failed to show expected improvement with age, and even masked true group differences (Connolly and Grantham-McGregor 1993; Baddeley, Gardener et al. 1995; Oluyomi and Houser 2002). The inadequacy of the adoption approach results from the fact that activities used to measure psychological concepts reflect values, knowledge and communication strategies of their culture of origin. This fact is acknowledged by those publishing psychological tests that have been rigorously standardised on large-scale populations, culminating in revisions of their publications designed for other population groups. One example is the British Ability Scales, modified to become the Differential Ability Scales for use in the USA (Collins, Smith et al. 1990). How much more caution then should we employ in transferring tests between cultures with more obvious differences between contexts?

The two other approaches, Adaptation and Assembly, both acknowledge the need to account for specific cultural influences. Assembly is the production of a totally novel test, based upon the cultural practices of the target population, and makes no assumptions about conceptual or performance comparability. It can be applied to avoid the construct bias that occurs when existing instruments only partially sample the domains that define a

construct. For instance, Western tests of intelligence emphasize skills such as reasoning, memory and acquired knowledge, but lack the social component of African conceptualisations of intelligence. Therefore, while sub-tests of the K-ABC (Kaufman and Kaufman 1983) or WPPIS-R (Wechsler 1989) may provide an adequate appraisal of specific cognitive skills, they do not provide an appropriate definition of intelligence in the African context. Assembly is most appropriate where there is no already existing test to measure the concept being assessed, or where an underlying psychological concept is most readily observed through an activity that is culturally specific (Kearins 1976; Sternberg, Nokes et al. 2001).

Adaptation, in contrast, can be followed when an existing instrument provides a proven measure of an underlying psychological concept, but where the specific methodology used in one context (test language, materials, and/or administration procedures), requires modifications to make it suitable to the new context (Foxcroft 2002; Holding, Taylor et al. 2004). Adaptation acknowledges the existence of underlying psychological universals, and attempts to enable the measurement of cognitive or behavioral skills in a universally comparable manner. During the process of translating and adapting an instrument, Herdman's (1998) universalist model suggests aiming for the following forms of equivalence: conceptual, item, semantic, operational, measurement and functional, to ensure that the comparability between similar tests adapted for different contexts is maintained. Only if the last two, measurement and functional equivalence, are achieved will it be possible to compare performance scores between test versions. While being time-consuming, establishing equivalence ensures that the adaptation maintains acceptable reliability and validity and can provide meaningful interpretations of test scores. However, any changes will mean changes to the initial standardization, and we would argue that even if equivalence is established, each test version should be supported by its own standardisation and normative population.

The choice of whether to adapt or to assemble will therefore largely depend upon the psychological construct of interest. Some psychological constructs show greater functional universality than others. One example is psychomotor development, where motor control and co-ordination are seen to develop in a universal sequence. Despite this, assessment items and normative tables will vary from one setting to another to take into

account not only the different rates of development seen across settings (Neil 1972; Leiderman, Babu et al. 1973; Lynn 1998; de Vries 1999) but also the different activities with which children are familiar. For example, in most rural settings in Africa, children do not have stairs in their homes; therefore items that assess psychomotor development based on the ability to climb stairs may be inappropriate. Other psychological phenomena, such as parenting practices, are more obviously defined and influenced by the specific cultural milieu. Consequently, measures of parenting behaviour may not readily transfer, and may call for the development of measures based on local definitions of appropriate parenting behaviour.

The advantage of Adaptation over Assembly is the degree of comparability that it affords between study sites. When seeking to understand the influence of specific health exposures, information on common constructs that can be used to summarise effects will be invaluable. With this in mind we have, in the main, selected the Adaptation approach. Through extensive experience in the field we have developed the following systematic procedure to ensure: that the tests used are measuring what they are purported to measure, and that the results are meaningful to the population in which the tests have been used. The procedure can be divided into four main stages, with equivalence of the adapted instrument being evaluated according to at least one of each of the equivalences suggested in Herdman's model (Herdman 1998) (listed above).

The Kilifi 4-stage approach to test adaptation

STEP 1- *Construct definition:* In this first stage, the aim is to clearly define the construct to be measured, as well as the cultural parameters that will delimit the definition of the concepts involved. Multiple sources should be used to generate a description of constructs in ways that are both culture-specific (emic) as well as in culturally neutral terms (etic), that will facilitate universal comparisons across cultures (Pike 1967). Activities undertaken will aim at establishing definitions of the parameters of interest. This includes identifying: specific activities that will define variation in skill levels in the function of interest, a conceptual vocabulary through which the concept can be described in the language of interest, and finally, the training needs of those who will describe and administer the assessment. This information can be collected through:

1. A **systematic review** of existing literature, to provide a critical evaluation of existing measures of the construct of interest, and help identify a suitable instrument for adaptation;
2. The use of a **Panel of Experts,** to develop a conceptual vocabulary. A culturally appropriate definition of the construct of interest can be developed by a panel of people who can contribute to at least one of the following areas of expertise: psychological (with a background in, for example, community mobilisation, child development and behaviour change); cultural (members of the target community), linguistic (fluent in or at least familiar with the language of the proposed test takers). We have, for example, run workshops with groups of community nurses and field workers involved in health research. Through discussion and activities aimed at developing lay assessments of the concept of interest, they then prepare *a glossary of relevant terms* in the language of the target population. We have found producing a glossary of terms prior to exposure to the original instrument important in training non- experts in how to undertake conceptual rather than literal translations. The production of conceptual translations is important so that the sensitivity, content and face validity of an instrument are not compromised.
3. **Community participation** While professional panels will include community members, we have observed that the process of training in modern techniques, such as biomedical medicine and research, exerts a significant influence on the understanding of the new concepts being introduced. To adequately infer understanding at a community level of proposed assessment tasks and questionnaire items, it is also important to consult lay members of the community whose exposure to education mirrors the spread of that of the majority of the population. Methods for eliciting community understanding include Focus Group Discussions (FGDs), individual in-depth interviews, and direct observation, outlined in some detail by Abubakar, van de Vijver et al. (2008). The role of the target community is diverse and may include: definition of the construct in the local community, providing face validity of the item/ tasks and evaluating the cultural appropriateness of the items/tasks (Abubakar, van de Vijver et al. 2008).

STEP 2 - Item pool creation: The aim of this step is to prepare a list of potentially acceptable items, in a clear and unambiguous language, through the integration of the information collected in Step 1. Using this information on cultural practices and available vocabulary, original items from existing instruments are vetted for their appropriateness. No item should be discarded until it has been rigorously evaluated, as this can lead to premature removal. An example is, during the assessment of the home environment of rural children, our initial assumption was that an item on exposure to television viewing would be irrelevant. However, this item later proved to be sensitive to between household differences, as children had access to televisions within the neighbourhood, even though it was only a few that had one in their own houses. At this stage too, additional items from those that are felt to reflect local behaviours are added to the item pool to provide potential substitutes for discarded items.

To prepare materials in the appropriate language, the World Health Organisation (WHO) recommends the use of a single schedule translated by a bilingual panel comprising members with related competencies. In practice, we have found the translation procedure recommended by the WHO (2007) to have fundamental flaws, many of which are also outlined by Leplège and Verdier (1995). One limitation is the assumption that an adequate vocabulary exists in the target language. Another is that the audience of the new language version will be familiar with the concepts in the original document. In addition the WHO translation process also assumes that a translation team that has expertise in the target concepts and are highly skilled linguists can be assembled. The reality of both a lack of relevant expertise and budgetary constraints means that available personnel are often native speakers of the target language, but they are neither professional translator, nor are they necessarily familiar with the topics to be investigated. For all these reasons, we have followed a translation system similar to that described by Gandek, Ware et al. (1998). We begin with the initial translation of the schedule using the glossary of relevant terms as a guide, to produce a conceptual rather than a literal translation. This version is then evaluated for semantic and conceptual clarity through the comparison of multiple back translations. The steps in the evaluative process can be summarised as: 1) Production of at least two back-translations by two independent native speakers; 2) Evaluation of conceptual equivalence by a study panel through comparison

of the multiple translations and, 3) Production of a second draft incorporating modifications to replace problematic items or response choices. This draft is then given to a second set of translators for back translation. The process continues until there are no more semantic differences between translations, showing that the essential meaning of the items has been understood. In the process, those items that remain poorly understood will be dropped.

STEP 3 – *Developing the procedure*: The main aim is, through pre-piloting, to produce a schedule of items of acceptable length, as well as clear guidelines for their administration. An iterative process is used, with each version or sub-set of items trialled on 5-10 participants. Items are discarded if they produce little variation in response, and administrative procedures are modified if they elicit negative reactions from participants. One such example is the community reaction to a task evaluating the development of self-recognition in which infants were required to look into a mirror, an activity that is taboo in several African societies. Administration techniques, such as one-to-one interaction between the assessor and the child, have also been observed to reduce response variation in several African communities (Harkness and Super 1981). Table 1 outlines the principles that should govern the evaluation of the test adaptation process, and summarises the methods by which that evaluation can be carried out. A schedule of items and administrative procedures is then drawn up for piloting.

(Insert table 1)

STEP 4 – *Evaluation of adapted schedule*: The aim of this step is to establish the basic psychometric properties of the adapted instrument. To enable a detailed evaluation, the schedule drawn up through Step 3 should be administered to at least 75 participants. Standard psychometric evaluations, outlined in Table 2, are used to determine a final schedule of items and the overall appropriateness of the instrument (Anastasia 1988).

(Insert table 2)

Training and the Production of Manuals and Guidelines[1]:

We have already referred to the probable lack of experience and prior training of assessment staff. Tasks and tests that are complex to administer and depend upon extensive previous training may not be suitable to the African context, and administrative manuals should be developed that are also piloted for clarity of language and procedures. We have developed a curriculum to ensure adequate preparation of an assessment team, and estimate that its delivery requires a minimum of 3-4 weeks of instruction, practice and evaluation, for a team with no previous testing experience to achieve a minimum acceptable standard.

The curriculum is divided into two parts. Part one introduces broad issues related to assessment in the context of developmental psychology; topics covered in this section include theories of child development, basic research methods, data collection techniques and ethical issues in research. This provides a conceptual background found to be essential to the understanding the rigours of standardised assessment procedures. In the second part, we impart specific skills related to using the measures to be administered, and evaluate test administration performance according to a structured performance schedule that sets a minimum level of competence.

Issues of Interpretation of Results

Our experience shows that, through following rigorous procedures, one can adequately adapt measures for use in Africa from those initially developed in the West (Holding, Taylor et al. 2004; Abubakar, van de Vijver et al. 2008; Alcock, Holding et al. 2008) .The potential pitfalls of relying merely on translations have been outlined before. However, it should also be acknowledged that any changes, even minimal translations, bring into question the applicability of the initial standardisation, and the possibility of the inappropriate use of standardisation tables, leading to mis-interpretations and misdiagnoses (Losen, Orfield et al. 2002). An adequate control group will overcome many issues of interpretation and analysis within a new context. Between contexts, statistical techniques, such as effect sizes, enable us to make cross-cultural comparisons in the absence of directly comparable raw scores.

[1] Readers interested in accessing the manuals, guides and psychological measures we have developed should contact the first author or write to admin-amhf@africaonline.co.ke

To promote the availability of rigorously produced measures, and partially circumvent the time consuming process of test adaptation, we strongly advocate that researchers and clinicians working in Africa share their data on test performance. By combining data across multiple contexts we can identify the cultural boundaries of a test, and build up a much needed test library of appropriate assessments.

References

Abubakar, A., A. J. R. van de Vijver, et al. (2008). "Monitoring Psychomotor Development in a Resource-Limited Setting: An Evaluation of the Kilifi Developmental Inventory." Annals of Tropical Paediatrics **28** 217-226.

Abubakar, A., F. van de Vijver, et al. (2008). Enhancing the Validity of Psychological Assessment in Sub-Saharan Africa through Participant Consultation. Selected Papers from the 18th International Congress of the Association for Cross-Cultural Psychology IACCP.

Alcock, K., P. A. Holding, et al. (2008). "Constructing test of cognitive abilities for schooled and unschooled children." Journal of Cross Cultural Psychology.

Anastasia, A. (1988). Psychological testing. New York, Macmillan Publishing.

Baddeley, A., J. M. Gardener, et al. (1995). "Cross-cultural cognition: developing tests for developing countries." Applied Cognitive Psychology **9**: S173-S195.

Carter, J. A., J. A. Less, et al. (2005). "Issues in the development of cross-cultural assessments of speech and language for children."

Collins, E. D., P. Smith, et al. (1990). British Abilities Scales London, Harcourt Assessment

Connolly, K. and S. Grantham-McGregor (1993). "Key issues in generating a psychological-testing protocol." American Journal of Clinical Nutrition **57**: 317-318.

de Vries, M. (1999). "Babies, brains and culture: optimizing neurodevelopment on the savanna." Acta Paediatrica **Suppl, 88**: 43-48.

Foxcroft, C. D., Ed. (2002). Ethical issues related to psychological testing in Africa; What I have learnt so far. Online readings in psychology and culture. Washington, Western Washington University.

Gandek, B., J. E. Ware, et al. (1998). "Cross-validation of item selection and scoring for the SF-12 Health Survey in nine countries: results from the IQOLA Project. International Quality of Life Assessment." Journal of Clinical Epidemiology **51**: 1171-8.

Harkness, S. and C. M. Super, Eds. (1981). Why African children are so hard to test? Cross-cultural research at issue. New York, Academic Press.

Herdman, M., Fox-Rushby, J, & Badia, X. (1998). "A model of equivalence in the cultural adaptation of HRQoL instruments: the universalist approach." Quality of Life Research **7**: 323-335.

Holding, P. A. and P. Kitsao-Wekulo (under review). "'Is assessing participation in daily activities a suitable approach for measuring the impact of disease on child development in African children?' "

Holding, P. A., H. G. Taylor, et al. (2004). "Assessing cognitive outcomes in a rural African population: Development of a neuropsychological battery in Kilifi District, Kenya." Journal of the International Neuropsychological Society **10**: 246-260.

Jukes, M. C., M. Pinder, et al. (2006). "Long-Term Impact of Malaria Chemoprophylaxis on Cognitive Abilities and Educational Attainment: Follow-Up of a Controlled Trial." PLoS Clin Trials **1**: e19.

Kathuria, R. and R. Serpell (1998). "Standardization of the Panga Munthu Test- A nonverbal cognitive test developed in Zambia." Journal of Negro Education **67**: 228-241.

Kaufman, A. S. and N. L. Kaufman (1983). "Kaufman Assessment Battery for Children: Administration and scoring manual." Circle Pines, MN, American Guidance Service.

Kearins, J., Ed. (1976). Skills of desert children Aboriginal cognition: retrospect and prospect. Canberra, Australian Institute of Aboriginal Study.

Leiderman, H. P., B. Babu, et al. (1973). "African infant precocity and some social influences during the first year." Nature **242**: 247-249.

Leplège, A. and A. Verdier (1995). The adaptation of health status measures: methodological aspects of the

translation procedure. The international assessment of health-related quality of life: Theory,

translation, measurement and analysis. S. Shumaker and R. Berzon. Oxford, Rapid communications of Oxford.

Losen, D. J., G. Orfield, et al. (2002). Racial inequity in special education. Cambridge, MA, Civil Rights Project at Harvard University, Harvard Education Press.

Lynn, R. (1998). "New data on black infant precocity." Personality and Individual Differences **25**: 801-804.

Neil, W. (1972). "African infant precocity." Psychological Bulletin **78**(5): 353-367.

Nell, V. (2000). Cross-cultural neurospsychological assessment: theory and practice. Mawhah, NJ, Lawrence Erlbaum Associate.

Oluyomi, A. O. and R. F. Houser (2002). "Yoruba toddler's engagement in errands and cognitive performance on the Yoruba Mental Subscale." The International Society for the Study of Behavioural Development **26**: 145-153.

Pike, K. L. (1967). Language in relation to a unified theory of structure of human behaviour. The Hague: Mouton.

Reynolds, C. R. (1983). "Test bias: In God we trust: All others must have data." Journal of Special Education **17**: 242-260.

Scarr, S. (1984). Race, Social Class and Individual Differences in IQ. Hillsdale, New Jersey, Lawrence Erlbaum Associates.

Serpell, R. (1979). "How specific are perceptual skills? A cross-cultural study of pattern reproduction." British Journal of Psychology **70**: 365-80.

Sternberg, R. J., C. Nokes, et al. (2001). "The relationship between academic and practical intelligence: A case study in Kenya." Intelligence **29**: 401-418.

Wechsler, D. (1989). Wechsler Preschool and Primary Scales of Intelligence-Revised New York, The Psychological Corporation.

WHO.(2007).http://www.who.int/substance_abuse/research_tools/translation/en/index.htmlProcess of translation and adaptation of instruments".

*Table **1.** Guidelines for Item Selection*

Principle	Methods of Evaluation
• Relevance to the construct • Relevance to the community • Clarity of language being used. • Clarity of instructions • Acceptability of the chosen method of administration • Suitability of the testing environment	• Item score variance, • Participant feedback and community knowledge • Multiple translation process • Test session observations, • Feedback from administrators • Error analysis, • Correlation of responses with other assessments/measurements

Table 2: A summary of statistical considerations in test evaluation

Psychometric consideration	Description	Statistical Technique/Recommended cut offs
Item Level Analysis		
Item variability	Distribution of item scores to look for floor, ceiling effects and overall distribution or responses	Descriptive Statistics- No one response selected in excess of 75% (or if screening for unusual behaviours 90%)
Development and Evaluation of Summary Scores		
Internal reliability	Intercorrelation of items within a test	Cronbach's alpha; split half reliabilities
Test-retest reliability	Correlation of measures between two time-points	Intra class correlations (Consistency- a more robust approach than *r*)
Inter-tester reliability	Correlation of measures taken by 2 assessors	Intra class correlations (total agreement - a more robust approach than Kappa)
Inter-form reliability	Evaluating equivalence of two item schedules	Correlation analysis of scores from the 2 forms
Concurrent validity (including criterion validity)	Relationship between test under construction and alternative measures of same concept (e.g. current best practice) taken simultaneously	Correlation between the scores from the 2 tests
Convergent validity	Relationship between abilities theorised to be closely related	Correlation between measures of closely related skills e.g. measures of Language and Verbal IQ.
Divergent validity	Lack of relationship between abilities theorised to be unrelated	Lack of correlation or lower correlation measures of 2 different skills e.g. measures of IQ and motor skills