Test Transfer and Application

A Structure for Managing and Evaluating the Re-Standardization Process of Tests of Cognition, Child Development and Family Influences

Prepared by the International Centre for Behavioural Studies

for the Grand Challenges Canada Saving Brains Program

Project funded by:
Grand Challenges Canada™
Grands Défis Canada™

Contents

Introduction

Purpose

The purpose of this document is to help structure the process of adaptation and standardization of tests by:

- providing a practical guide for **the preparation of tests and questionnaires** where there is no normative data or previous evidence of reliability and validity for test materials;
- offering a method to **record and evaluate test measurements** to aid in the analysis of individual studies, and to make comparisons between studies;
- giving guidelines for the **systematic collection** of data on test adaptation and evaluation; and
- **supporting** those with a limited background in psychometrics.

Evaluation Focus

The evaluations of tests need to focus on the following psychometric aspects of test performance:

1. **Variability** in scores between individuals, the underlying principle being true population variance or the ability of an instrument to measure real differences between respondents.
2. **Reliability**, providing evidence of consistency:

   a) between the items making up the test, *the internal consistency*;

   b) of scores over time, *the test-retest reliability*;

   c) between people making the assessment, *the inter-rater reliability*.
3. **Validity,** providing evidence that the test/questionnaire *is* measuring what it *intends* to measure. There are multiple aspects to validity so, given the lack of appropriate Gold Standards, we will focus on:

   a) what the test means to the target population, *face validity*;

   b) the relationship between measurements of similar and different performances, *convergent/discriminant validity*;

   c) any additional variability explained by the target test, relative to that provided by other tests, *incremental validity*.

   d) for tests derived from those previously published and standardised on other population groups another important aspect to examine is the theory that underlies that test, as articulated in the original manual:

   - thus any variability in performance as outlined in the original manual, such as differences by age or gender, should be investigated in the new data set;
   - the relationship between sub-tests (*Construct Validity*) should also be explored and compared to the original test framework.

Definition of Terms

Concept ..... The underlying idea, meaning, or trait (the underlying variable of interest, e.g. physical growth)

Construct ..... The aspect of the underlying variable actually to be measured (e.g. height)

Content ..... What the items look like and the material used (e.g. pictures of football players)

Full–scale score ..... Combining all, or a significant number, of test scores within the battery (e.g. combining *to compute* and *IQ/mental processing*)

Item ..... The parts of the test, or the constituent elements of the construct used to measure/estimate the construct (e.g. a measure of length at 3 months or standing height at 3 years) or a single component of the test (e.g. a question or an individual task)

Procedure ..... How the test is administered (e.g. self-fill questionnaire, naturalistic observation, type of instruction)

Sub-scale score ..... Combining sub-groups of tests (e.g. to compute a memory scale, or a fine/gross motor scale)

Standard Error of Measurement (SEM) ..... The error expected of an individual's test score

Sensitivity ..... A measure of identification error (do individuals with some impairment score high on a test but are incorrectly identified as non-impaired)

Specificity ..... A measure of identification error (are individuals without impairments identified as non-impaired)

Test ..... The instrument used to measure a construct (e.g. height meter, questionnaire or performance measure)

Test battery ..... Group of tests used to measure a general concept; batteries may provide multiple measures, including individual test scores, full-scale scores and sub-scale scores

Developing Your Test Battery – Four Steps

To develop your battery of tests you should take four steps, which can be applied even if you have already selected specific tests.

To complete these steps efficiently it is worth being supported by:

- **A review panel**, consisting of people whose combined expertise covers the following:
    - knowledge of psychological concepts (or the concept of interest);
    - experience in assessment, particularly of children;
    - fluency in the languages *spoken* (not written or academic) by the target population.

Aim for a team of 8-10 people if possible, to address not only differences in expertise, but also to obtain a more random sample of knowledge, opinion and experience.

- **A translation team** to carry out forward translations and back translations. While the initial translation should be completed by somebody with a good knowledge of both languages involved in the process, back translations are best carried out by people fluent in the 'street' language used by your target population (see p.8).

The four steps are explained briefly, follow the page link for the full procedure.

Step 1: Concept Definition (see p.6)

This is a *preparation step* to provide a working glossary of terms and phrases in the target language to help those who are going to translate the items, as well as identifying relevant skills and behaviours salient to the target population. The glossary is especially necessary for questionnaires, where a vocabulary of common psychological terms may not exist in the language in which the test will be used. With a pre-prepared vocabulary items can be evaluated as possible/impossible to translate, and potential replacement items can be suggested by the concepts or constructs identified as relevant to the target population.

Step 2: Item Pool Creation (see p.8)

In this step original items are evaluated for cultural appropriateness, adjusted if necessary, and eliminated if they fail to engage the target population.

Step 3: Developing a Procedure (see p.11)

Aspects to be considered are the appropriateness of: the medium (e.g. drawings vs. photographs); the test location (e.g. home vs. school vs. clinic); the instructions (e.g. written vs. oral, number of sample/practice items); and so on as well as the impact that changes make to test score variability.

Step 4: Psychometric Evaluations (see p.13)

This step provides an evidence base for how the chosen tests perform with regard to consistency and concept definition. That is test reliability and validity.

The following pages go into each step in depth and suggest both a procedure to follow and a way to summarize the output from each step. The summary, in turn, provides a way to evaluate the performance of each test.

Step 1: Concept Definition

Aim: to develop a vocabulary and a framework to judge the relevance and appropriateness of the test.

Selecting the Test

For each concept/construct/test that you want to use ask each of the questions in the first column and see how the answers fit against the checks in the second column.

Question	Check
What is it that the study wants to know?	Are you covering relevant functional skills?
How can I best express the concept of interest in the language of the study population?	Is a vocabulary available? Is the construct, so expressed, associated with that concept within the target population?
Is there a pre-existing test that might be a good estimate of the concept of interest?	Are you unnecessarily re-inventing wheels?
What construct is this selected test measuring? How is it measuring it?	Is the methodology appropriate to your situation? Does it match: time, cost, skill level of the staff, and content and procedural familiarity of the population?
Is there any evidence for the validity and reliability of the construct in the population where you intend to apply the test?	Are there any gaps in your knowledge of how the test(s) operates in your population group?

Summarise the Information

Use the following table, or something similar, to record and monitor your activities so that you can track any changes made and describe the stages of test development you have completed. When you come to publish your material this is the information reviewers are increasingly interested in. They want to see the detail and the rigor with which the modifications were made.

Concept	I. As a panel discuss and record the means of expressing the concept in: a) the local language b) simple English	II. List related constructs that are valued or understood in the local population in: a) the local language b) simple English	III. List specific questions or pictures that are relevant to the constructs agreed under II.
e.g. General intelligence			

Step 2: Item Pool Creation

Aim: to choose appropriate questions and pictures and put them in the correct format (language and material).

The Pool

Items can be drawn from multiple sources:

- If you have begun with an instrument that already exists as your core, initially retain ALL the original items in the first round of panel review, and try as far as possible to translate them.
- Supplement from other tests, or suggestions derived from constructs that are locally relevant to your target population.
- In a summary table list **all** the items available for consideration, and that will be pre-piloted.

Tracking Changes

It is useful to summarise all the changes you make, from the initial stages to the completion of the study. With this information you can then evaluate the following aspects of the modified test in relation to the original:

1. Changes in content.
2. Changes in order, as they relate to item difficulty or fluency.
3. Through back translation, how the changes relate to the original intention, and the feasibility of maintaining equivalence.
4. Background characteristics of those being assessed or interviewed. How does education, SES, gender, age, etc. affect responses or performance levels?

The spreadsheet Tracking Changes v4.xls provides a useful format to collect this data in one database. Open Tracking Changes

Translations

Making a conceptual translation of items

Begin with one core translation from the original language (foreign to the target population). It is not necessary to make several translations in the first stage of development, especially if the translator is well-briefed in the conceptual framework of the material. If you have access to a second expert translator, this translation can be back translated to the original foreign language to provide a comparability to the original translation.

The key to evaluating how 'equivalent' the translation is, is to have at least three back translations made by a minimum of 2-3 people NOT previously briefed on the material and who are familiar with **local usage** of the target language. A comparison of multiple

versions will help the panel ensure that the material is unambiguous and the interpretation of the meaning is likely to be consistent across respondents.

As far as possible, use representatives of the target population since it is their understanding that is crucial. These back translations will give a picture of how the words are understood in the new language and the process may need to be repeated several times for some items. For each round of back translations use different people, or the impression the new material makes will be contaminated by familiarity with the old material.

Issues to consider in the *translation/back translation* process

Back translation is designed to ensure equivalence between the translated item (the new item) and the original version.

There are four 'equivalences' that should be maintained, as far as possible:

1. Conceptual ..... the new item needs to tap into the same conceptual idea

2. Item ..... the item should be representative of the same level of difficulty, or contribute in an equivalent manner to the total score

3. Semantic..... the new item should, where possible mean the same thing (this can conflict with 1 & 2 above).

4. Procedural..... the new item should, where possible, be administered in an equivalent manner to the original.

Therefore back translations need to look at BOTH the content of questions and the instructions to the administrator, to support an equivalence of procedure.

The back-translations should be compared by a review panel (see above). Where the versions agree with each other, and with the original intention, conceptually, they can be accepted.

If there is some disagreement at a semantic level between at least two of the translations then look for a more precise term. If all three disagree then a re-translation is needed, and the translator will need to be better briefed as to the meaning being sought.

The changed sections need to go for *re-back translation*, to a different pool of people if possible, until the whole document is understood in an equivalent manner by the 'audience' (who are your back translators at this point).

Be aware of the educational level of your eventual target audience. They are the ones who need to understand the material, not an academic population. Avoid the use of high level grammar and vocabulary.

Pre-pilot

Pre-pilot the items initially selected on a small number from the target population (at least 5), to verify the clarity of the initial list. Each subsequent change should also be tried out on about 2-5 respondents. The panel should review all this information before

selecting the schedule for the main pilot. Keep a record of who is in your pilot group, and your outcomes.

Evaluate

The main aim of this phase is to reject items that are not appropriate to the target group. Items may be rejected:

- If there is no local vocabulary or image that defines the item.
- For items that will contribute to a summary score if the item shows no, or very restricted, variance in the target population (e.g. more than 95% can/not answer, do/not have).

The exception to this second rule for rejection is item(s) that show little variability, but whose presence or absence is strongly indicative of risk or resilience. This includes items such as food availability and abnormal behaviours. This is an issue of discriminant power of an item that also needs conceptual evidence.

Evaluate your final schedule using the following:

Main Selection Criteria	**Methods of Evaluation**
• Acceptability of the chosen measure	• Participant feedback • Test session observations
• Relevance to the community • Clarity of language being used	• Review of the translation process
• Clarity of instructions	• Feedback from administrators • Review of performance on first few items to establish need for extended sample items
• Relevance to the construct	• Correlation of test results with performance on other assessments/measures • Comparison to community beliefs
• Sensitivity	• Item score variance
• Suitability of the method of administration (Presentation of the items, place of administration & type of explanations required)	• Error analysis: Review the performance of each item by using % responses are: Incorrect Deliberately not administered (e.g. too difficult) Child refused Missed in error by the assessors
• Conduct of assessors	• Review item order using % of correct responses • Differences between total scores by assessor

Step 3: Developing a Procedure

Aim: to make sure the instructions and the scoring are understood by the administrator and the person being interviewed or tested.

During piloting you need to address how the items are presented to reduce the error variance associated with unfamiliarity with testing.

Potential Hurdles and Beneficial Strategies

- If a child is not familiar with being tested they may need more time or more practice items may need to be introduced.
- If the population is not familiar with questionnaires, or cannot read fluently, then questions have to be asked orally in a confidential but clear manner.
- If the assessors are new to assessing, then scoring can be a problem. Score sheets and instructions need to be simplified and regularly checked by someone familiar with the assessment procedure.

In addition to establishing whether the items themselves work, important questions to address during piloting are:

1. How much practice or prompting is needed to understand the purpose of the test or the questions?
2. How long is needed to carry out the procedure?
3. Is the scoring clear/unambiguous to the administrator?
4. What is the correct order of difficulty for this population?
5. Has the right balance of items and tests been selected?

Questionnaire Issues

If you are going to use a questionnaire in a population with limited experience of questionnaires, and with limited literacy, then to maximise its ability to elicit information will require you to use these three guiding principles:

- oral presentation
- discourse format
- adequately validated vocabulary

The application of a system of measurement that is alien to the respondents will significantly affect the reliability and validity of the measurement process. It is not appropriate to expect respondents to learn the expectations of the questionnaire format in the course of one interview. You should adapt the process to suit the respondents by modifying the interview content and the manner of presentation.

There are certain **basic principles of questionnaire design that increase the reliability** and sensitivity of measurement capabilities:

- use language that is sufficiently clear to stimulate an informed response; and

- describe the concepts clearly enough that even small differences between subjects can be established.

The **major hurdles** to achieving these goals in many of our settings are:

- The lack of a subtle enough vocabulary sufficient for ordinary people to understand differences between closely related behaviours.
- The need to provide examples to ensure that the questionnaire is understood, which lengthens the procedure.
- The lack of experience or exposure to concepts/constructs amongst the interviewers asking the questions, which is associated with:
    - difficulty in providing accurate examples to help the respondent understand the questions; and
    - limitations in their ability to consistently interpret the descriptions given by respondents in relation to the concepts/constructs being explored.

Furthermore the standard interview format requires respondents to:

- Provide direct responses to questions on information personal to the family.
- Share information with a stranger.
- Describe behaviours in terms of discrete constructs, rather than integrated behaviours that relate to understood and accepted behavioural 'types'.

The social constructs involved in the question and answer format of questionnaires are foreign to many cultural settings, which restricts the ability to engage respondents in the interview procedure. Experience has demonstrated that to address and overcome the hurdles described above interviews should be conducted in a format that respects the expectations, understanding and beliefs of the respondents. Hence the three guiding principles listed at the beginning of this section.

Step 4: Psychometric Evaluations

Aim: To establish the psychometric properties of a test in a new context.

Logistics of a Pilot Study

A test with proven reliability and validity in one context may not maintain those properties in another context, due to differences in exposures and values. It is thus highly recommended, to ensure that you can properly interpret your data, to CALIBRATE your instruments to the new context in which they are to be applied.

Most texts on psychometric evaluation recommend that a minimum sample of 75–100 per cell of interest should provide acceptable reliability. This is not always feasible prior to a study.

A compromise approach could be:

1. Ensure that each person who is to administer the tests collects data on a minimum of 5 respondents on the full schedule.
2. A minimum target of 20 should pick up issues of clarity. As numbers increase over this threshold the data provided will have increased stability.
3. Stratify your sampling to include children from across the target age range and to adequately represent your target population.
4. Debrief the team regularly to check on issues related to vocabulary, scoring and interpretation of behaviour/performance. Record these qualitative observations, for later checking against error analysis of actual scores achieved (see point 5).
5. Review the measurement properties of the tests in this small group to look for any specific problems. (% no responses, % correct/incorrect responses, the ordering of item difficulty).
6. Continue to collect data as the study progresses, and carry out an interim analysis when you reach over 75–100.
7. Where you are intending to apply the test(s) to measure the effect of an intervention or risk exposure include a control group that reflects a random sample of the general population. You can use your control group to examine psychometric issues post hoc. The danger is that you might find, after the event, that the test was not reliable, but at least you can then interpret the data you have in the light of this knowledge.

Throughout the pilot study you should address whether you have the evidence to maintain or change:

1. The content of individual Items (words/pictures)
2. Item order
3. Instructions

To answer those questions you also need to be able to answer the following:

1. What is the evidence for test score variability, or variability in responses? For example evaluate scores in relation to the normality/skewness of the distribution.
2. How representative of the final population was the pilot sample?

Psychometric Summary

There are many ways to examine the psychometric properties of the test you are using. Several are defined below.

Psychometric consideration	Description	Statistical Technique; Recommended Guidelines for Evaluation
Item Level Analysis		
Item difficulty	Distribution of item scores to look for floor, ceiling effects and overall distribution or responses	No. of test-takers who answer the item correctly/incorrectly/fail to respond. Calculated on the basis of the relationship of correct to incorrect responses using linear and non-linear regression techniques
Internal Consistency	To describe the contribution of individual items to a total score. Contribution is reduced when responses are selected, or fail to be selected, by < 75% (or if screening for unusual behaviours 90%) of respondents.	Inter-correlation of items within a test, also using Cronbach's alpha; split half reliabilities, ICC as appropriate Item-test: To what degree an item score and the test score measure the same thing? Item-rest: To what degree an item-score and test score without the item, measure the same thing?
Item discrimination	Calculated to separate medium-level performers from the best/worst ones through correlational analysis Distractor-test: What proportions pick the distractor, say 0 when the correct answer is 1?	Correlation coefficients are computed using the usual formula for product-moment or Pearson correlation.
Graphical Item Analysis	A simple way of presenting differences between any pair of populations (e.g., boys/girls)	The item difficulty (p-values) are plotted against each item in Excel or SPSS

Evaluation of Summary Scores

Psychometric consideration	Description	Statistical Technique; Recommended Guidelines for Evaluation
Test-retest reliability	Consistency and stability over time, usually measured at two time points, 3-4 weeks apart	Correlation co-efficient (*r*) Intra class correlations (a more robust approach than *r* especially when sample size is low (i.e. lower than 15).
Inter-rater 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 administered close together but in random or reversed order.	Correlation between the scores of 2 forms of the same test
Face Validity	To explore the appropriateness of test content to a social context	Descriptive and qualitative analysis
Concurrent validity (including)	Relationship between test and alternative measures of same concept (e.g. current best practice) measured at the same time	Correlation between the scores from the 2 tests
Criterion validity	Similarly to above, but measured at a different time (later)	Correlation between the scores from the 2 tests
Convergent validity	Relationship between constructs/abilities theorized 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 constructs/abilities theorized to be unrelated	Lack of correlation or lower correlation measures of 2 different skills (e.g. measures of IQ and motor skills)
Construct validity	Provides evidence of test structure and the relationship between constructs measured	Confirmatory or Exploratory Factor Analysis
Estimate of true performance	Provides evidence on the accuracy with which an individual's score approximates the true score.	Standard Error of Measurement (SEM)

Implications for data collection, entry and analysis

To carry out the proposed evaluations each study will need to:

1. Enter the data for tests that are constructed of multiple items at the item level, at least for the first 100 individuals. This data will then allow you to look at other aspects of item analysis for consistency.
2. Organise the testing schedule to enable a regular schedule of two raters/administrators looking at the same performance. This means **either** that two raters observe the same assessment **or** that a careful record is made of output (by filming, keeping drawings, a transcription of verbal responses) and two people rate or code the same output.
3. Assess a group of children (>50 preferably >75) on two separate occasions, on the same **performance measures**, after an interval of 3-4 weeks. To control for other sources of variance the same assessor should carry out both the test and the re-test.

Estimating True Score Variance

A person's **true score** is the mean score they would have achieved if given the test an infinite number of times. Using multiple measurements would reduce the variability in performance that is due to unsystematic changes that effect performance of the same test taker at different times, or of different test takers (random error).

While the true score will never be known, it can be estimated by comparing an observed score to the level of consistency/stability (reliability) of the test. That is, true score variance can be estimated by using the ratio of a measure of test reliability and the observed variance in test scores in a population.

The range, the variance, and the standard deviation across the whole sample will be indicators of true score variance if the sample is representative and sufficient in size.

Maintaining standardisation

There are a number of approaches.

a) Carrying out observations of test administration (can double up with inter-observer evaluation);
b) Daily review of data forms and feedback to assessors;
c) Comparison between assessors of their data forms and scoring techniques;
d) Over the course of the study you should try to target that 10% of assessments are evaluated through the calculation of inter-observer reliability.

Close supervision is the key. Identify a supervisor, or gold standard assessor, who has a regular schedule of observations built into his/her schedule.

Suggested Key Indicators of Test Performance

Evaluation	Issues to consider
Estimates of True score variance using:	Requires a representative sample of the larger target population
Internal consistency reliability	Requires data entry of item scores
Test-retest reliability	Requires measurement of the same test at two time points
Inter-rater reliability	Requires at least two observations of the same data by separate raters
Maintenance of standardization over time	Requires repeated observations of assessments, monitoring about 10% of sample over the study
Sensitivity to the effects of background characteristics	Requires data on contextual characteristics that may independently influence variability in outcome
Construct validity	Requires data on at least 10 observations (children) per test included in the factor analysis

General Information

Contact Information

International Centre for Behavioural Studies

incentbs@gmail.com

www.icbstudies.org

Grand Challenges Canada Saving Brains Program

savingbrains@grandchallenges.ca

www.grandchallenges.ca

Useful Links

Summary of Saving Brains projects: www.grandchallenges.ca/savingbrains-grantees-en/

Tracking changes spreadsheet { HYPERLINK "C:\\My Documents\\MyExcelFile.xls" }

References

The framework used is derived from the approach described in:

Where there are no tests: A Systematic Approach to Test Adaptation (2010) Holding P, Abubakar A, & Kitsao Wekulo P. Chapter 9 In: Cognitive Impairment: Causes, Diagnosis and Treatments *Ed*: Landow, ML Nova Science Publishers, Inc. Series: Neurology - Laboratory and Clinical Research Developments.

This article expands on some of these concepts, e.g. issues of back translation:

Is assessing participation in daily activities a suitable approach for measuring the impact of disease on child development in African children? (2009) Holding P, Kitsao-Wekulo P, Journal of Child and Adolescent Mental Health 1728-0591, Volume 21, Issue 2, Pages 127 – 138 http://www.tandfonline.com/doi/abs/10.2989/JCAMH.2009.21.2.4.1012

Other texts discussing related issues are:

Alcock, K. J., Holding, P. A., Mung'ala-Odera, V., & Newton, C. R. J. C. (2008). Constructing tests of cognitive abilities for schooled and unschooled children. *Journal of Cross-Cultural Psychology, 39*, 529-551.

Anastasi, A; (1997). *Psychological Testing* (Seventh ed.). Upper Saddle River (NJ): Prentice Hall. ISBN 978-0-02-303085-7.

Ardila, A. (2005). Cultural values underlying psychometric cognitive testing. *Neuropsychology Review, 15*, 185-195.

Berry, J.W., Poortinga, Y.H., & Pandey, J. (eds.) (1997). *Handbook of cross-cultural psychology: Vol. 1: Theory and Method* (2nd ed.). Boston, MA: Allyn and Bacon.

Berry, J.W., Poortinga, Y.H., Segall, M.H.& Dasen, P.R. (2002). *Cross-cultural psychology: research and applications.* New York: Cambridge University Press. 4

Connolly, K. J., & Grantham-McGregor, S. M. (1993). Key issues in generating a psychological-testing protocol. *American Journal of Clinical Nutrition, 57*(Suppl. 2), 317S-318S.

Geisinger, K. F. (1994). Cross-cultural normative assessment: Translation and adaptation issues influencing the normative interpretation of assessment instruments. *Psychological Assessment, 6*, 304-312

Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist, 52*, 1115-1124.

Hambleton, R. K. (1994). Guidelines for adapting educational and psychological tests: A progress report. *European Journal of Psychological Assessment*, 10, 229-244.

Hambleton, R. K., & Kanjee, A. (1995). Increasing the validity of cross-cultural assessments: Use of improved methods for test adaptation. *European Journal of Psychological Assessment* 11, 147-157.

Hambleton, R.K. (2006). Chapter 1: Issues, Designs, and Technical Guidelines for Adapting Tests Into Multiple Languages and Cultures. In: Hambleton, R.K., Merenda, P.F.,

& Spielberger, C.D. (2005/6). *Adapting educational and psychological tests for cross-cultural assessment.* Lawrence Erlbaum Associates: Mahwah NJ.

Harkness, S., & Super, C. (1977). Why African children are so hard to test. *Annals of the New York Academy of Sciences, 285*, 326-331.

Holding, P. A., Taylor, H. G., Kazungu, S. D., Mkala, T., Gona, J., Mwamuye, B., Stevenson, J. (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. H., & Grigorenko, E. L. (2010). Assessment of cognitive abilities in multiethnic countries: The case of the Wolof and Mandinka in the Gambia. *British Journal of Educational Psychology, 80,* 77-97.

Kawabata, M., Mallett, C. J., & Jackson, S. A. (2008). The flow state scale-2 and dispositional flow scale-2: Examination of factorial validity and reliability for Japanese adults. *Psychology for Sports and Exercise, 9,* 465-485.

Kline, P., (1993) *The Handbook of Psychological Testing*, Routledge: London and New York.

Malda, M., van de Vijver, F. J. R., Srinivasan, K., Transler, C., & Sukumar, P. (2010). Traveling with cognitive tests: Testing the validity of a KABC-II adaptation in India. *Assessment, 17*, 107-115.

Malda, M., van de Vijver, F. J. R., Srinivasan, K., Transler, C., Sukumar, P., & Rao, K. (2008). Adapting a cognitive test for a different culture: An illustration of qualitative procedures. *Psychology Science Quarterly, 50*, 451-468

Nampijja, M., Apule, B., Lule, S., Akurut, H., Muhangi, L., Elliott, A. M., . . . Alcock, K. J. (2010). Adaptation of western measures of cognition for assessing five-year-old semi-urban Ugandan children. *British Journal of Educational Psychology, 80* (Pt 1), 15-30.

Poortinga, Y.H., & Van der Flier, H. (1988). The meaning of item bias in ability tests. IN S.H. Irvine & J.W. Berry (Eds.), *Human abilities in cultural context* (pp. 166-183). New York: Cambridge University Press.

Ruffieux, N., Njamnshi, A. K., Mayer, E., Sztajzel, R., Kengne, A., Ngamaleu, R. N., . . . Hauert, C. A. (2010). Neuropsychology in Cameroon: First normative data for cognitive tests among school-aged children. *Child Neuropsychology, 15*, 1-19.

Solarsh, B., & Alant, E. (2006). The challenge of cross-cultural assessment: The test of ability to explain for Zulu-speaking children. *Journal of Communication Disorders, 39*, 109-138.10 *Assessment XX(X)*

Van de Vijver, F.J.R., & Leung, K. (1997a). Methods of data analysis and comparative research. In Berry, J.W., Poortinga, Y.H., & Pandey, J. (eds.). Theory and method (pp257-300). *Handbook of cross-cultural psychology: Vol. 1: Theory and Method* (2nd ed.). Boston, MA: Allyn and Bacon.

Van de Vijver, F.J.R., & Leung, K. (1997b). Methods and data analysis for cross-cultural research. Newbury Park, CA: Sage. 10. Van de Vijver, F.J.R & Tanzer, N.K. (1997). Bias

and equivalence in Cross-cultural assessment: an overview. *European Review of Applied Psychology*, vol. 47, 263-279.