Assessing the impact of early interventions upon human capital formation

Exploring the Costs and Benefits of Test Adaptation

This report represents the hours of time and effort that Saving Brains teams have expended on preparing measures of child performance. It will focus on the practical lessons of this process, lessons that have implications for producing reliable indices, so needed to detail the relationship between early intervention and later Human Capital Formation

"An important part (of preparation)…is determining how the system will be installed and what factors can affect its operation once it's up and running in your process line "

http://www.automation.com/library/articles-white-papers/process-weighing/five-factors-that-can-affect-your-weighing-systems-accuracy

To ensure the process of measurement is accurate we have to consider both the instrument of measurement as well as the system into which these instruments are being inserted……

The Challenge

Metrics of Development are key to providing the evidence to guide the promotion of human capital formation. The development of the Saving Brains programme has enabled a detailed exploration across multiple settings of the capacity of applied metrics to define and evaluate the impact of interventions in the crucial early years. In this report on the contribution of a metrics framework, the Core Metrics from Round 1 - Re-Enrolment Studies, we examine the process of preparation, administration as well as the measurement properties of the tests applied, and make recommendations for the preparation of future Metrics frameworks.

The operational difficulties of applying a single testing protocol across multiple sites, as described in the experience of many of the Saving Brains projects, undoubtedly reflects to some degree the effect of cultural/ social/linguistic influences (the context) upon the manifestation of underlying skills and abilities (brain function). However the challenges experienced are not uniform. More detailed evidence that identifies aspects of test functioning that require contextualisation to accurately measure impact would aid in addressing these challenges.

P.A.Holding 2014

CONTENTS **Page**

Key Issues and Practical Implications

Point 1: The evidence here acquired demonstrates that adequate test properties can be maintained even when aspects of the tests have been altered to account for contextual differences. In fact, for the first time it was possible to demonstrate that modifications were, in many instances, associated with improved reliability. *The pattern of results suggests the need for more careful item selection, and perhaps lengthier piloting of content changes, while changes to administration procedures were relatively easy to make, with clearly beneficial effects.*

Point 2: Although "no responses" were identified as a reason for making modifications to tests, there was no systematic investigation of whether specific sub-groups were disproportionately represented in the missing data. A recent publication from Zambia supports the importance of evaluating the ability of an instrument to represent achievement and change in the whole population, as an added but separate assessment of test properties. *The evaluation missing data, seldom reported, should become a core part of the reporting procedure.*

Point 3: As most studies were still in the preliminary stages of analysis, there was insufficient data to establish the validity of the measurement metrics applied. However, given the lack of opportunity to track populations over time, the predictive validity of any instrument is rarely explored in human development literature. *One more logistically feasible process towards filling this gap would be to have a greater number of before and after designs, enabling the measurement of stability and consistency following interventions, rather than the more common cross-sectional approach of most study designs.*

Point 4: The cost of using tests in studies of the size of the re-enrolment studies (i.e. samples in the thousands) reached upwards of USD 100- 200 000. The main cost appears to have been not the tests themselves, but the payments publishers require for each individual record form, a recurrent cost. *Should publishers wish their tests to contribute to impact evaluation they need to recognize the difference between population based research, which includes no personal gain, and clinical applications, where the assessment team may use the material to elicit a fee.*

Point 5: The time taken to prepare test time was not an insignificant component of each study, a process that should begin after ethical approval has been received to engage children and their families. *Ethical review committees need therefore to recognize the need for piloting, and to be flexible in their requirements for reviewing final instrument details and record forms.*

Point 6: The cost of applying tests includes not only time and money, but also, in the narratives of some teams, the stress of preparing measures within a narrow time frame. Many reported that they had not previously been aware how complicated the process would be. This comment reflects the complexity of measuring human capital, and the limitations of the current resources available. *The establishment of an open access resource, into which the rich experience of the Saving Brains groups can be deposited, to be withdrawn and updated by subsequent rounds would be a major contribution to reducing time, cost and stress.*

Point 7: Without data on validity and responsiveness it is not possible to draw conclusions on the value of the Core Metrics themselves. It was possible to find reliable instruments for each. As yet the meaning of the Core Metrics in understanding Human Capital Formation remains under-explored. *A summary analysis across all studies, examining the pattern of results, and the size of the effects observed, would provide a more detailed guide of the value of these test, these Metrics, and help identify the focus for future studies.*

Background - Exploring the Rationale

Metrics selected to define meaning, strength of association, and the influence of time and place facilitate a detailed description of the process of growth and development following early interventions. What defines an effective process of measurement is a key focus interest and debate amongst the research and implementation communities.

The assessment of intellectual capacity, enshrined in the Intelligence Quotient (IQ test), has long been a source of intellectual and academic deliberation. The controversy has focussed upon the relative strength of *nature* (capacity is fixed or rigid) versus *nurture* (where capacity is malleable, and an individual's developmental trajectory can be altered through stimulation or education). Programmes such as Saving Brains reflect the acknowledgement that this debate is not *Either/Or*. Developmental trajectories are modifiable, through the reduction of risk and or by stimulating resilience, with the size of the effect varying, influenced by age, length and intensity of exposure to factors conveying risk and resilience, as well as according to the behaviour, skill or function under consideration. Ensuring both clarity and accuracy of measurement is therefore crucial to unravelling the complexity and relative size of these influences upon human development.

Another pertinent debate in developmental psychology is that of *universality versus cultural specificity* in the development of thinking, reasoning and functioning. Resolution of this debate directly impacts upon the development of a theory of change generalizable to working in multiple settings across multiple cultures. While universality suggests a common structure and developmental trajectory underlying the acquisition of skills and abilities (which can therefore be tracked using a single set of measurement tools), cultural specificity suggests significant limitations to the accuracy of data collected by measures that do not address linguistic and cultural diversity. The core question is not whether, but how much contextualisation is necessary.

Most approaches to measuring neuropsychological functioning developed in "western" settings demonstrate a degree of within population variance in other cultural settings, supportive of the notion of universal applicability of measurement concepts. The variance can, however, be very restricted, thus limiting the responsiveness of measurement tools to even major stressors, and consequently also the development of an adequate theory of change. Even in the early stages of development the order of acquisition of different skills varies between cultures as closely related as that of Europe and North America (Vierhaus et al 2011). Other manifestations of cultural specificity are differences in the expressions of psychological concepts as well as differences in the triggers of common behaviours. An example of the latter; laughter is universally recognisable, but its manifestation is not always associated with joy. An example of the former; a capacity for memory storage and retrieval is universal, although different cultures stimulate a greater capacity in either auditory or spatial working memory (Chione, & Buggie, 1993, Wagner 1974 Kearins, 1986).

Language, Culture, Thought and Biology

Literature has shown a universal relationship between published tests of general intelligence and academic achievement, albeit that the strength of this association varies by socio-cultural group.

The effect of culture etc. on test performance has been well documented in the literature, from as long ago as the studies of Vygotsky and Luria in the 1920's and 30's. There have also been several seminal studies that suggest that while there can be the appearance of statistical equivalence, even minor differences in true score variance can lead to misrepresentation of effects and effect sizes. Statistical evidence for the bias in the calculation and interpretation of results that can follow, even for children from minority groups within a single catchment area, is also well documented (Alcock et al). But these biases will not stem purely from potential differences in the way information is perceived or processed.

Assessment is an interactive process. Between the question and the answer of any assessment protocol lies a social space, in which influences upon engagement, understanding and motivation also act upon how a test will be completed. If these, even subtle, differences are ignored, and it is assumed that identical measures can be used identically across multiple sites to quantify the same cognitive functions, with the same level of confidence, potential limitations in analysis and interpretation may result.

Epigenetics, in recognising and describing the combined influence of genetic predispositions, and environmental activation, has stimulated the beginnings of a paradigm shift in the conceptualisation of the measurement of risk and resilience, and of the continuity of human potentials. Achievement and wellbeing is conceived as deriving from the interaction between common (universal) underlying structures and abilities upon which unique (context specific) influences act to create diversity. Expanding the field of enquiry to include experiences in and of the majority world will therefore bring increasing clarity on what we should be measuring, and how we should be measuring it (ref African child).

Toolbox or Toolkit?

The assumption that one test (or battery of tests) is an appropriate vehicle to enable an adequate, and accurate, evaluation of impact is based upon several assumptions:

a) The existence of core universals in cognition, brain function and development.

b) Human behaviour can be adequately summarised by selecting a restricted range of key components of behaviour

c) The importance of these key components is equally recognised and shared across multiple contexts.

d) That the measurement properties of a measurement instrument are directly transferable across time and place.

The more universal the manifestation of concepts and constructs the greater the possibility that a toolbox of uniform tests will be relevant to multiple sites. The greater the specific the influence of culture or exposure to risk, the more appropriate is the idea of a toolkit that provides a contextualised framework to guide assessment. Given the as yet limited data from majority world settings to substantiate or refute these assumptions, understanding the complexity of human behaviour, and the brain behaviour link in the face of multiple risks, continues to be an under-explored frontier of scientific inquiry. The increasing body of evidence supports an interaction between universal cognitive predispositions, with task-specific constraints on the manifestations of underlying abilities. Advances in statistical and methodological approaches will enable us to move the frontiers even further forward.

The Saving Brains programme provides a unique opportunity to explore and reflect upon the process of test application in a number of different settings. It not only provides data on performance and development in the majority world, but also potentially the means to explore methodologies that enable meaningful comparisons and summaries across settings. Moving tests across and between populations necessarily requires addressing the language spoken by the test taker. A systematic evaluation of the impact of other modifications made upon the performance of the tests themselves will begin to address the important question of the accuracy of the measurement rubrics being deployed.

Definitions of Test Equivalence

For a test to be accepted as "appropriate' in different contexts there needs to be proof of cross-cultural equivalence. There are multiple layers of equivalence, which have different implications for analysis and interpretation of responses/ scores on a test. Each form of equivalence does not guarantee the presence of another form, and they may, in certain circumstances, be mutually exclusive.

The different forms of equivalence can be defined as**:**

1. **Conceptual**: the test needs to tap into the same conceptual idea as originally intended. This is the highest order of equivalence, and can be demonstrated through common patterns of association between different tests, and through error analysis, demonstrating similar processes applied by different populations in the solution of the test or problem presented.

2**. Item**: Each item, and the combination of items, should be representative of the same level of difficulty as originally intended, or contribute in an equivalent manner to the total score. Item difficulty analysis would illustrate whether the same scoring rubric is valid across contexts, with the distribution of scores indicating common or unique relationships of scores indicative of the same diagnostic categories in different contexts.

3. **Semantic**: Items as presented should maintain the same intention as originally envisioned. Evidence of semantic compatibility is the easiest to observe, but can be the most difficult to maintain across languages. Finding appropriate substitutions, that also maintain item equivalence, is a major challenge, and the reason why there is reticence in some quarters to make adaptations to individual items.

4. **Procedural**: The administration procedures should elicit an equivalent response as the original intention. An equivalent administration over time and place, the principle that lies at the heart of a standardized test, should ensure an "equal playing field" for all test takers. But let us not forget the primary intention of the design of the process of administration, which is to support the communication of the test requirements, as well as to communicate the true attitudes or abilities of the respondent. If the administration process does not convey with equal clarity across time and place the intention of the test, then the value of a single standardized process comes into question. The proportion of no responses, limitations on test –re-test reliability, and differences in distribution of scores are all potential indicators of a lack of procedural equivalence, even if the actual methodology applied appears to be the same.

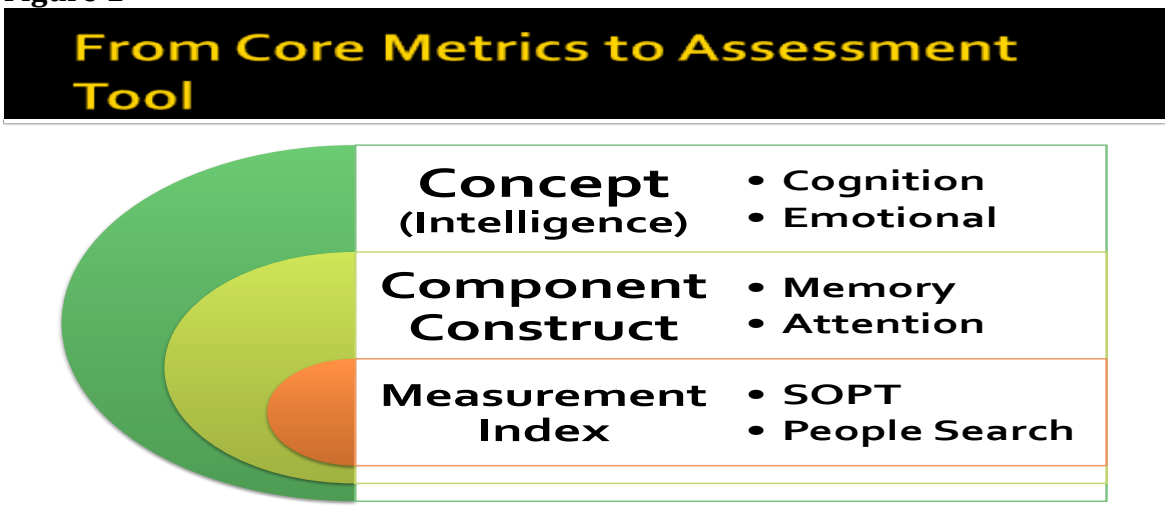
Evidence of a lack of equivalence on any of these levels may require test adaptations to be made to ensure that both **conceptual** or **measurement equivalence** in maintained. Adaptations may therefore be necessary to support fidelity to the original intention of the test.

The concept behind the CORE METRICS

Among the potential approaches to measuring cognitive development across multiple sites is a *one size fits all approach* predicated upon the concept of *universality*. While ostensibly providing a means of summarising across sites, a single Index (or tool) ignores the potential limitations that inherent differences between populations in the acquisition and manifestations of skills, however small, might have on the ability to combine data across sites. A difference in the underlying structure of performance scores (differences of scale) brings into question the accuracy of a simple consolidation of data into a common pool. A single index may also restrict the scope of the understanding of human capital formation to that described by a narrow range of potential functions (or constructs).

The Core Metrics approach aims rather at identifying common concepts rather than a common constructs or content, thus, hopefully, overcoming both limitations of scale and scope. Figure 1 illustrates the differential specificity described at the level of Concept, Construct or Index. Where an Index, a specific tool, provides the means of estimating the levels of a latent variable, the Construct, that describes more specific components of broader psychological Concepts.

Figure 1



Despite evidence of statistical bias, there is still an expectation that published tests, having undergone rigorous development, will provide more robust measures of human capital formation than contextualised measures. Altering the content of such tests may certainly undermine the robustness of the original index. However, as the evidence of test reliability and validity is usually only available from test application in populations with very diverse experience to the new target group, assumptions of that robustness lacks evidence, and needs to proven in alternative settings. By not assuming universality, but by systematically collecting data on test equivalence and test performance, the underlying logic of measurement will be better and more widely recognized, as will the understanding

of the relationship between the social and the biological components of performance on tests, potentially stimulating theoretical advances in the understanding of human capital formation.

The application of the Core metrics in the Saving Brains Programme

Saving Brains is a programme whose core focus is to determine positive and efficacious strategies that improve the developmental status of vulnerable children. The objective is to identify interventions that will scale up, by identifying activities that have a measurable and long-term impact.

In its first round eleven independent studies, situated in 14 different sites, across low and middle-income countries shared a common core of assessment metrics. The Core Metrics were selected through consensus amongst study groups and external consultants, to identify concepts considered key to establishing a positive impact upon development. The Core Metrics encompassed multiple constructs, demonstrated through a broad range of behaviours. Agreement was also reached in the initial discussions amongst the re-enrolment teams over some key constructs. For example working memory and attention shift were identified as important constructs within the concept – Executive Function. The selection of the instruments to measure these constructs, the actual indices of performance, the tests themselves, was left up to the discretion of the separate teams.

The intention of the core metrics was to build a framework that would bring in a level of consistency and comparability across the multiple studies involved in the programme. The opportunity was therefore provided to examine the process of test adaptation, explore the relationship between aspects of the adaptation process and accuracy in measuring impact, and to summarize the value added of a specific indices or measures.

The initial round of RE-Enrolment studies (RE) covered children aged from 4 years of age, until young adulthood (20 years of age). Subsequent rounds of the programme, Scaling Impact (SI) have modified this list of Core Metrics, to take into account the different stages of development of the target children in the funded studies.

Core Outcome Metrics - RE

1. Height for Age
2. Years of Schooling
3. ***Estimate of General Intelligence***
4. ***Measures of Executive Function***
5. ***Indications of Literacy***
6. ***Presence of Behavioral and Emotional Problems***

Core Outcome Metrics SI

1. ***Cognition***
2. ***Gross and Fine Motor***
3. ***Expressive and Receptive Language***
4. ***Social Emotional Capacity***

Test Adaptation – The Surveys

The exploration of the test batteries was conducted using responses to two surveys, supplemented with data provided by individual RE study groups that reported on the reliability and validity of the instruments applied.

Survey 1 - **Core Metrics General Information** - concerned background details of the specific context(s) of assessment, a general description of the assessments considered, as well as those that were excluded.

The purpose of collecting this information was to form the foundation for a test bibliography, as a key resource in identifying useful indices of human capital formation. The background details provided information on key features of the environment in which the teams were working that might influence preparation time, cost and the availability of resources to support the implementation of an assessment protocol.

Information elicited included:

- Location where the tests were applied (country and regions)
- Languages of those regions
- Educational Background of the general population
- Experience of assessment teams
- Instruments applied
- Instruments excluded

Survey 2 – Instrumentation - was designed to collect data at the test level on details of the adaptation process, as well as information on test application that can also be used to guide future test selection.

Information elicited included:

- The instrument source
- The cost of preparation
- The time taken to prepare the material
- Summary of any modifications made
- Normative or comparison groups available
- Evaluation of the instrument as an index of development

The results of the surveys, and of the psychometric data shared, will be reported to

A. Describe the material currently available upon which to develop a Test Bibliography
B. Describe the Challenges Faced to Test Preparation
C. Explore the relationship between test psychometric properties and changes made to address contextualisation.

Developing a Test Bibliography

Infrastructure

i. Location: Country and Language

The survey data was provided by nine of the 11 study groups from the Re-Enrolment Grantees (RE). In addition we also have information from Scaling Impact Grantees (SI), addressing their initial selection procedures of assessment tools.

Test materials and instructions were prepared in a variety of different languages. Amongst the SI teams the study populations spoke 11 different languages, with only 3 of the 9 study groups working in a single language.

Continent	Country	Language
Africa	Ghana	Twee
		Sena
	Kenya	Kiswahili
		Dholuo
	Malawi	Chichewa
		Sena
	South Africa	Zulu
		Xhosa
		English
	Tanzania	Kiswahili
	Zambia	Tonga
Asia	Bangladesh	Chittagonian Dialect
		Bengali
	India	Urdu
	Indonesia	Indonesian
		Sasak
	Pakistan	Sindi
		Awadh
	Vietnam	Vietnamese
South America	Columbia	Spanish
	Peru	Spanish
		Quechua

ii. Location: Population Structure (a) Urbanisation and (b) Underlying Literacy levels

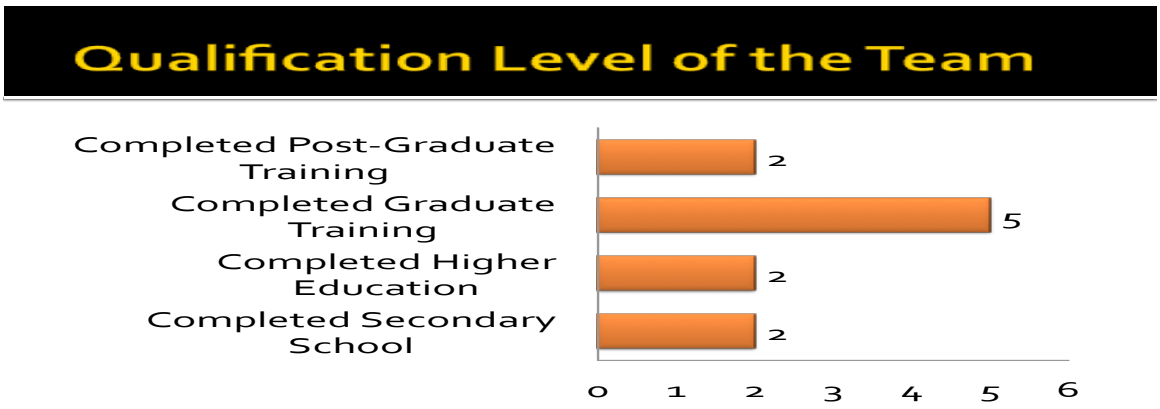
(a) The majority of the study sites were in rural communities.

SB Round	Rural	Peri-urban	Urban
RE	60%	20%	20%
SI	50%	25%	25%

(b) As an index of the literacy levels of the general population, groups were asked about the Primary School Enrolment of the regions in which they worked. In the RE group in only one area, in South Africa, was primary attendance reported at 100% coverage. While 50% of study groups reported that the majority of the children (defined as > 70%) attend school, three groups reported primary school attendance of less than 50%. In these latter groups, these rates were below the national average for that population, suggesting particular needs of these targeted populations. Amongst the SI teams only two study sites were reported to have universal schooling, even at primary level.

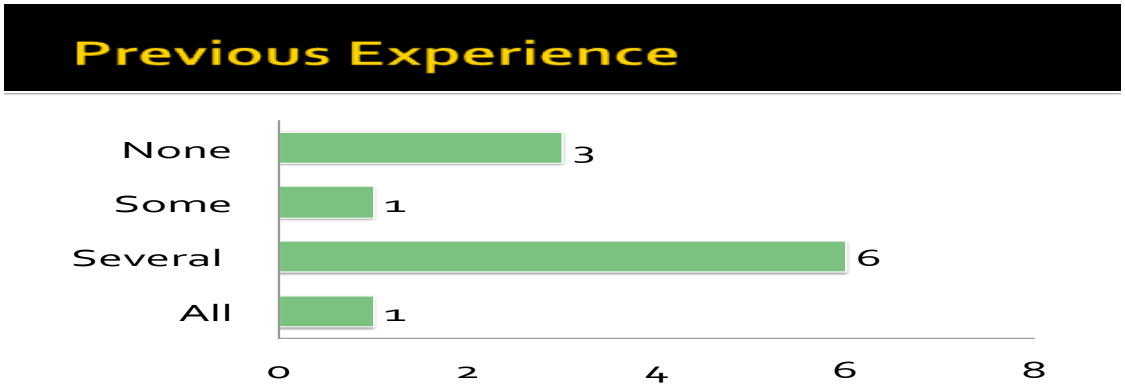
iii. Child Assessment Infrastructure

Personnel: Assessment Teams were recruited from those who had completed, in the main, higher education. Although several RE groups were able to recruit graduates, most teams were without extensive assessment experience.



Amongst the SI teams there was a lower level of formal education, and again, as with the RE teams, illustrated below, with limited previous assessment experience (8/9 SI groups had no or limited previous experience).

Previous assessment experience reported for RE assessment teams.



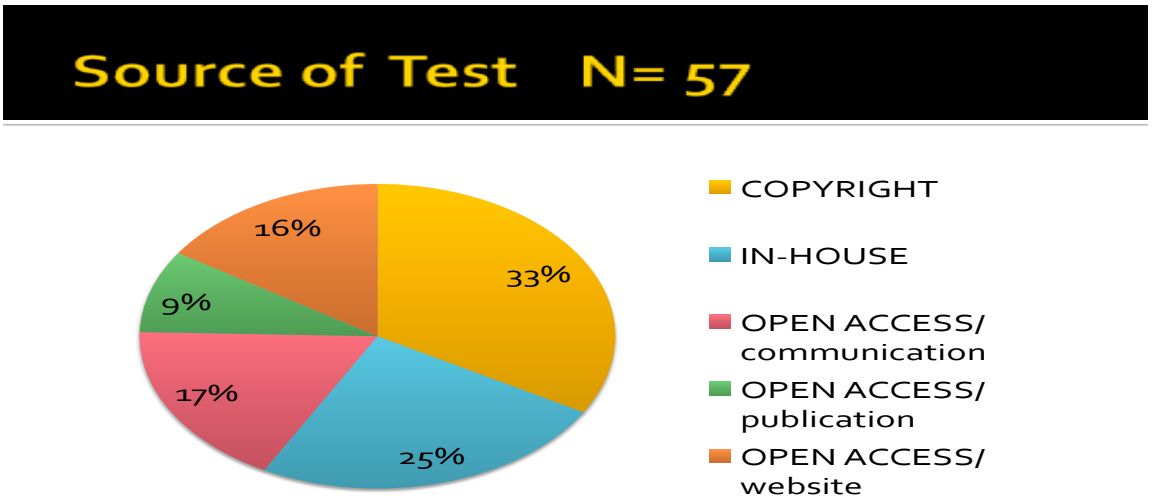
Location of Assessment: Only two study groups had access to a dedicated assessment centre, others used clinics, schools or homes.

Developing a Test Bibliography

Test Selection

The instruments that will be reported on were applied by 9 study groups, working in 12 different sites, and included data on the application of 57 tests or test batteries.

The RE studies reported on batteries that included between 5 – 29 tests. The source of these tests was reported as follows. Many teams used, as far as possible, test materials that were already available for the context, with a quarter of all tests reported upon developed in-house.



Tests selected, and later applied, are listed in Appendix 1. Those initially selected but later excluded, with the reasons given for that exclusion are identified below.

Excluded tests

From the NEPSY - A Developmental NEuroPSYchological Assessment

Sub-test	**Reason Given for Exclusion**
1. Attention	Replaced by another more suitable battery
	Problems with Colour recognition
2. Inhibition	Scoring Difficult for assessors
3. Block Design	Skewed distribution
4. Learning/Memory	Already covered in other tests

From the Kaufman Assessment Battery For Children II

Sub-test	**Reason Given for Exclusion**
5. Rebus	Low test-retest reliability; low population variance
6.	Needed too many changes in the time line
7. Verbal Knowledge	Needed too many changes in the time line
8. Expressive Vocabulary	Needed too many changes in the time line
9. Gestalt Closure	Needed too many changes in the time line

Others

Test –Sub-test	**Reason Given for Exclusion**
10. RNDA	Screening rather than diagnostic test
11. Movement Assessment Battery for Children	Replaced with a more familiar test
12. Snack Delay	No within population variance
13. Pick the Picture	Instructions not followed by young children
14. WAIS	Replaced by more suitable test
15. TEACh Score!	Instructions not followed by young children
16. Peabody Picture Vocabulary	Reason not given
17. D-KEFS - Card Sorting	Floor effects
18. Corsi Block Tapping	Instructions not followed by young children
19. Strengths and Difficulties Questionnaire	Rating Scale unsuitable for population
20. Zebra Lion- Inhibitory Control	Children did not like the puppets

Instrumentation

This section describes the process of test preparation. Information on preparation time, cost, and the types of changes made (the depth of adaptation) were considered in evaluating the 'value added' of test adaptation.

The Methodology Applied

The purpose of developing a measure of "depth of adaptation" is to provide a process through which to compare the time and cost of adaptation with the evidence of equivalence achieved. In other words, we are attempting to establish whether the process of adaptation followed, to maintain test equivalence, is associated with adequate psychometric properties of the test.

This task was an exploration of whether changes made to maintain in each one of the four levels of equivalence (conceptual, item, semantic, procedural) and address contextual challenges to data collection were associated with reliable and valid measures. To carry out this analysis we need to have data on tests from similar contexts on indices that were, or were not changed, and be able to examine evidence of within population variance, reliability and validity of the instruments.

Following an iterative process of consultation with 5 research sites, we determined that it was difficult to express "depth of adaptation" on a single dimension. Rather adaptations focus on multiple dimensions of tests. We have identified 4 key dimensions of change.

1. The Time and Cost of the Process followed
2. Changes to Content – Verbal and Images
3. Changes to Procedures – Materials and Administration
4. Changes to Calculating Appropriate Outcome Measures.

Whereby

2. **Changes to Content** follow-on from evidence of a lack of **semantic equivalence** across contexts and languages. (i.e. items cannot be used as original as children are not sufficiently familiar with the material to show within population variance).

   As evidenced from the location of the studies, test preparation will invariably begin with translation. Analysis will assume the ability to directly translate all content, and use data on changes in verbal targets that deviates from that premise.

3. **Changes to Administration Procedures** follow-on from evidence of a lack of **procedural equivalence** (i.e. children or administrators do not follow original instructions appropriately, and different procedures are needed in order for the appropriate responses to be elicited. An example here might be where children do not exhibit speed of completion on a timed task).

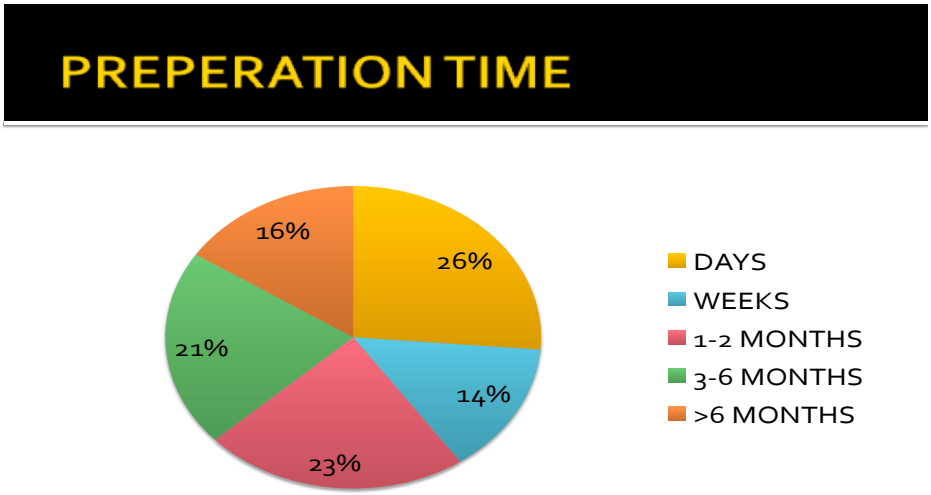
4. **Changes to Scoring Procedures.** The determination of appropriate scores follow from evidence of a lack of **item equivalence** (i.e. order of item difficulty or variance in scores).

The Details of the Adaptation Process

Time and Cost

Survey responses identify the need to plan for a minimum of 6 months of intensive and structured test and team preparation, when a battery of assessments is to be applied in a new context. The costs, both financial and in effort, also underscores the value of developing an accessible test repository of context ready measures to reduce the cost burden.

Time



Teams reported spending their time undertaking translations, piloting the new material. Those that reported spending greater than six months found in initial piloting very limited within population variance in test scores, and several iterations of new material were needed before a final version of new test material was ready for use. As illustrated by the following table, tests sourced from publishers (copyrighted) are likely to require as much preparation time as those being developed in-house.

The contextualisation of tests previously used in similar settings (largely available through open access) were observed to reduce the time required to initiate administration of a testing protocol.

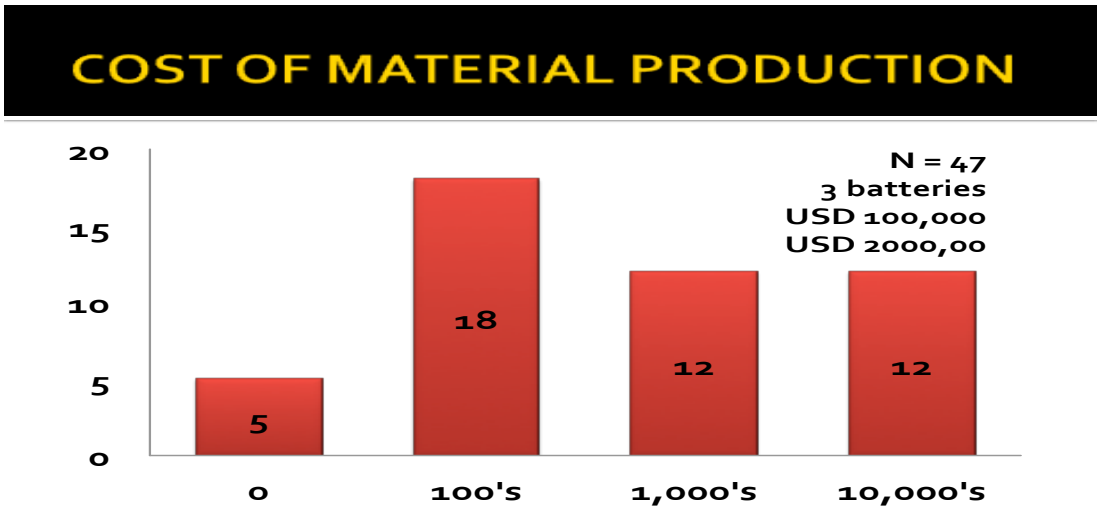
		% (N)			
SOURCE	**N**	**Days**	**Weeks**	**Few Months**	**Many Months**
Copyrighted	17	35 (6)	24 (4)	6 (1)	35 (6)
In House	6	33 (2)	17 (1)	17 (1)	33 (2)
Open Access	15	40 (6)	27 (4)	20 (3)	13 (2)

Costs

The benefit of a test repository, listing and providing access in multiple languages, is also illustrated by the cost implications of preparing tests for application. Estimated costs were supplied for 47 tests, ranging from free, to 10,00- dollars per test. Open access tests also significantly reduce the cost of the assessment process.

The estimated costs (in USD) of the material can be summarised as follows:

	Estimated Cost in USD		
SOURCE	**Mean**	**Min**	**Max**
Copyright	51,000	500	200000
In House	8,300	0	40000
Open Access	2,000	0	10,000



In the above table the number of tests reported at each cost level is reported on the y-axis, while the columns represented the estimated cost per test. A more detailed look at specific expenses, where people were able to provide a breakdown, is also informative for forward planning.

Expense	**Minimum in USD**	**Maximum In USD**
Adaptation (piloting)	500	5358
License Fee (including record forms)		76000
Production (e.g. art work)	100	500
Purchase	140	1240
Translation	100	2500

Of the twelve most expensive tests reported, half came from copyrighted material (cost per test was calculated from the total cost reported on the whole test battery). Much of this cost came from the price of the record sheets, which have to be bought per child.

For those using copyrighted material, the main cost is a recurrent one. Despite using a pre-published tool that incurs a large, and recurrent licensing fee, their narrative describing test preparation illustrates the multiple challenges and costs associated with fitting a test to the context.

> *Approximate estimation for our entire battery to be adapted is USD 200, 000. Approximately 9 months of work on adaption of the entire battery, requiring a team comprising 4 field assessors, one local psychologist, the PI. The Licensing fee for16 sub-test cost were approximately USD 76, 000, with additional expenses for and artist and a typist.*

The bulk of the costs of the six more expensive in-house tests were to cover programming the material and recording system onto a tablet. In other words, a one-off cost. Once the test is running on the tablet, the cost of the test is actually reduced, as there are no associated data entry costs.

Other specific examples shared by study groups provide further insights into future budgeting requirements.

1. Purchase of material: USD 1240. Cost of translation: USD 2000 Permission fee and cost of adaptation: USD 5358 Production of adapted material: USD 500 Total cost USD 9098

2. Total USD 3032.00 (Purchase of materials: USD 432, Translation: USD 2500, Production of additional materials: 100)

3. Cost of each iPad: $620 Cost of iPad app development: $20,000. Personnel/ expert consultation: $18,000

The details of the adaptation process

Changes to Content

Documented changes were made to:

a) Verbal Targets (spoken words, numbers)
b) Visual targets (photos, drawings, scripts)
c) Materials (the materials given the child to hold or manipulate eg pens sticks, cards)

a) Verbal Targets (spoken words, numbers)

Only 5 tests did not need translation, with 93% administered in a language other than English (4 in a mixture of local dialect and English), and 88% in a language different to the original source material.

The translation process was not straightforward, with a number of challenges facing those completing the translations. Once translated, only 14 of the 57 tests had no changes made in verbal targets. This means that less than half the tests were used without some degree of modification to the verbal targets. Of the 14 reported to have no changes, 5 were initially developed in-house, and the others were sourced in language versions suitable to the study population.

Challenges to the process of translation included:

- Children being **unfamiliar** with specific words.
- Local vocabulary only had a single word for **multiple concepts** included in the original items.
- The lack of a direct translation, requiring a **substitute in the same category**, and finally.
- The need to replace an **unfamiliar category** of words with one more familiar.

The following two tables quantify the extent of changes, and the reasons for those changes.

Proportion of verbal items retained as per the original

100%	**>50**	**< 50**	**All/1 or 2**
63%	9%	11%	17%

Types of changes

N = 37 (%(N))				No Changes
Unfamiliarity	Multiple concepts	Same Category	New Category	N = 57 (%(N))
56 (21)	14 (5)	19 (7)	11(4)	41 (23)

b) Visual targets (photos, drawings, scripts)

While images were changed less often, similar challenges to engagement of the children with the test material were experienced as with verbal items. Not all tests had images (18 were therefore excluded). Only 14 tests had images altered, with again unfamiliarity with the image presented being the most common reason.

Again the top table indicates the proportion of tests where changes were made, and the bottom the reasons given (similar to those for changes in verbal targets) for those changes.

Proportion of verbal items retained as per the original

100%	**>50**	**> 50**	**All/1 or 2**
64%	13%	8%	15%

Types of changes

N = 14 (%(N))				No Changes N = 39 (%(N))
Unfamiliarity	Multiple concepts	Same Category	New Category	
71 (10)		43 (6)	36 (5)	64 (25)

c) Materials (the materials given the child to hold or manipulate eg pens sticks, cards)

One study group replaced all the tests that were originally paper-based with a tablet-based administration and recording platform. Only three other tests had changes made to the medium of tests. One reported changing photos for drawings. Another changed the size of the print, and in a third case, an online test was taken off line.

Changes to Administration Procedures

Many more made some adjustments, even if only subtle, to the way the tests were administered (40%).

Examples of changes made were:

a. Expansion or modification of instructions to clarify meaning
b. Expansion or addition of sample or teaching items
c. Changes to time allowed the child to view material or to attempt the task
d. Other changes to the rules of the game
e. Changes in the context of test administration (how you are sitting, how the material is shared with the child, e.g the cards for story completion are not laid out, but handed as a pack to the child)

The most common alteration was to give children more sample items with which to familiarize themselves with the material and the instructions (22 tests). Longer time to view or practice items was accorded in 7 tests, while in 4 tests administration instructions were significantly changed, presumably to address the understanding of the children taking the tests.

Changes also included presenting what originally was written material, verbally. Although this was only actually reported on one test, this must have been more widespread in the administration of parental report questionnaires, given the educational background of the study populations.

Changes to Scoring Procedures

Sixteen tests had changes made to the scoring system. The modifications made included changes to:

a. Start point - The total number of items presented at each age range
b. End point – the rules for discontinuation
c. Changes to the application of timing to generate different scoring levels
d. Item order – to take into account differences in levels of ability
e. Changes to how the total score was calculated to provide a variation in scores within the target population

- In 5 tests changes were made following piloting to take account of observed differences in item difficulty order.
- 2 reported changing the start points for different ages, although none reported changing the discontinuation rules.
- The timing rules were changed for 2 tests
- The calculation of the full score was radically changed for 4 tests.

The majority of tests were to be analysed using the raw scores, as opposed to transposing the scores or using scaled scores from published materials. Only one study team reported that they intended to use published scaled scores, but given that items and item order had been modified it is assumed that they were no longer able to carry out that intention.

Association between adaptation and measurement characteristics

Six sites provided information on the psychometric properties of 41 individual tests. The information described the reliability of these measures. In the context of neurocognitive measurement reliability is synonymous with the concept of consistency, both in the variance across and between constituent items, as well as whether the measurement can be repeated over time.

Given the limited number of data points individual tests were grouped into their intended constructs, as described by survey responses, or according to the description of the tasks in related manuals. To summarize the data across studies the reliability statistics were further categorized according to the strength of the association, based upon accepted protocols and conventions that guide the evaluation of reliability statistics. (e.g. Cohen 1992).

Key: Strength of the Reliability Statistics

Excellent	Good	Limited	Poor	Very Poor
>/=.9	.7 - .89	.5 - .69	.3 - .49	< .3

Reliability Results

Studies reported upon the internal consistency, measuring the shared variance between constituent items, and the test –re test reliability, measuring consistency in test administration over time.

They also reported Inter – rater reliability, demonstrating consistency between members of the administration team. As a testament to the quality of the training received, the vast majority of tests achieved excellent consistency over different test administrators reported at or over .9.

The tables that follow display the association between constructs being measured and reported reliability. The tables are divided by the different Core Metrics, and by the type of reliability being examined. What is reported are the number of tests at each level of strength of association, with the most common level highlighted, as per the key displayed above.

Core Metric: Estimate of General Intelligence.

The levels of reliability achieved were consistent with the data provided by published test manuals. Seven-seven percent of tests of general intelligence achieved an internal consistency that can be considered good or excellent, with test-re-test reliability achieved by the majority of tests at the acceptable level.

Internal Consistency

Construct of General Intelligence	Excellent	Good	Limited	Poor	Very Poor
Sequential Memory		2	1	1	
Learning/Memory	2	2		1	
Planning/Perceptual Reasoning		2	2		
Simultaneous Processing	2	4	1		
Verbal Comprehension		2			
Vocabulary	3	1			

Test– Re Test

Construct of General Intelligence	Excellent	Good	Limited	Poor	Very Poor
Sequential Memory		3	2		1
Learning		1	4		
Perceptual Reasoning		6	1	1	
Simultaneous Processing	1	4	2		
Verbal Comprehension		2			
Vocabulary		4			

Core Metrics: Executive Function & Achievement (Literacy and Numeracy)

The reported reliabilities of the tests used to measure the core metrics of Executive Function and Achievement are a little more varied. There were fewer measures of executive function, with 60% achieving at least a "good" level of internal consistency reliability, suggesting overall that these adapted measures appear to be making reliable estimates of performance. However, the assessment of these higher order skills, the executive function, was less consistently reliable than those instruments measuring less complex cognitive skills. Again the levels of reliability were consistent with published data, although there was less of such material against which to compare the current data.

Construct of Executive Function	Internal Consistency Reliability				
	Excellent	Good	Limited	Poor	Very Poor
Working Memory		3			
Inhibition		1	1	1	
Attention-Shift		1			1
Attention Composite			1		
Cognitive Flexibility		1			

Construct of Executive Function	Test/Index	Test– Re Test				
		E	G	L	P	VP
Working Memory	Rey Ost. Recall/ HP/Digit Span Backwards		3	1		
Inhibition	Nogo			2	1	
Attention-Shift	Shift/Card Sort			2	2	
Sustained Attention	People Search/ Visual Search			1	1	
Attention Composite	DKH		1			
Cognitive Flexibility	CMS/Stroop		1	1	1	
Achievement						
Literacy/Numeracy			3		1	

The Validity of Measurements made

The key questions concerning test validity are whether the tests are measuring what they are supposed to be measuring (the underlying psychological constructs), and whether performance has a diagnostic ability, either concurrently or with regards to the prediction of future human capital formation (performance predictions).

Underlying Psychological Constructs are primarily measured through the association between the different tests or sub-tests.

Performance Predictions are measured through the association between test performance and external benchmarks.

This report was able only to draw upon discussions with different groups, who had only just begun this analysis. Preliminary analysis from four sites suggests that the tests are clearly estimating a strong general underlying factor. However, the association between the tests may not be as differentiated as anticipated by the a priori Core Metrics groupings.

Test Responsiveness

Responsiveness is an aspect of test quality that develops out of the brain – behaviour link. It has important implications for test selection, and the identification of appropriate comparison groups. It is measured by the association between key background characteristics and changes in test performance, and thus related to the sensitivity and specificity of an instrument to measure exposures of interest.

Instruments are commonly sensitive to:

1. Biological influences such as
    - Neurological Maturation – with performance increasing with age throughout childhood.
    - Nutritional Status – with performance decreasing in the face of chronic deficiencies.

2. Environmental influences such as
    - Schooling – This is not often considered in western settings, were there is compulsory school attendance. Previous literature from LMIC, where school attendance is less regular, clearly demonstrates the association between school exposure and test taking ability (Alcock et al, 2008).
    - Socio-Economic Status – where access to resources is associated with higher performance, an effect that increases with the age of the child. (Abubakar et al 2008).

3. Gender too has been observed to alter patterns of performance, either as a consequence of differences in socialisation or biology.

The associations reported came from ratings requested in the survey, using a 3-point scale, from highly related, through moderately related, to no association seen. For those who had not yet, or were not able to examine a specific association, a Do not know category was applied. The ratings are summarized in the table below, displaying the proportion and number of responses, at each level or rating, for each background characteristic.

These ratings suggest that environmental context exerts a measurable impact, larger than that estimated for general maturation, upon test performance. The implication being that these tests should be responsive to the impact of interventions designed to manipulate the environment. These ratings await the data to substantiate them.

	Highly	Moderately	No Effect	Don't know	Total # of tests
			38.10%	**61.90%**	
Age			8	13	21
		24.00%	**16.00%**	**60.00%**	
Gender		6	4	15	25
	12.00%	**28.00%**		**60.00%**	
SES	3	7		15	25
	4.00%	**20.00%**	**4.00%**	**72.00%**	
Schooling	1	5	1	18	25

The Investigation of the Added Value of Test Adaptation

The focus of this final section is a comparison of the test properties achieved following the different types of adaptations undertaken. Four test sites provided both details of individual test adaptation as well as the reliability by test, that enabled the comparison of adaptation with underlying measurement characteristics, primarily reliability, on 27 tests. These tests were divided into those that were adapted, and those that remained in there original form.

The following table summarizes the associations, which illustrate two main conclusions, that changes in procedures, in administration and scoring were associated with higher levels of reliability, while changes in content probably require a longer development time before the same benefits might be seen.

Adaptations to:	Alpha	Split Half	Test-re test	Adaptation Yes/No	N of Tests
Verbal Targets	0.67	**0.74**	**0.71**	YES	9
	0.82	0.69	0.57	NO	4
				NA	23
Visual Targets	**0.85**	**0.79**	0.61	YES	14
.	0.63	0.78	**0.65**	NO	15
				NA	7
Materials		**0.84**	0.60	YES	7
		0.74	**0.64**	NO	29
Administration	**0.77**	**0.78**	**0.71**	YES	23
	0.67	0.75	0.56	NO	7
				NA	6
Scoring	**0.76**	**0.81**	**0.72**	YES	16
	0.73	0.74	0.65	NO	14

While the data shows that changing the verbal content is associated with higher indices of split half reliability, as well as the improved consistency over time, this advantage was not seen in the alpha statistic.

The opposite trend was seen in the reliability properties following changes made to visual targets and materials, with a positive impact upon internal consistency, but none on consistency over time.

The changes made in administration, sometimes only subtle, can be seen to be associated with higher indices of reliability than the un-adapted tests. A similar effect is associated with changes in scoring systems.

Study Groups also showed distribution of test scores, not reproduced here, that indicated differences in underlying variance in score structures on the same test in different settings. Thus supporting the need to explore the methodology through which comparisons can be made across test administration sites.

Developing a Metrics Framework – International Considerations

Drawing Conclusions

The current expansion into communities with little or no previous exposure to psychological assessment has stimulated a debate on the how and the if of test application across different settings. The concern of the scientific community working within a cross-cultural paradigm is to ensure fidelity to the principles of the initial testing protocol. Under scrutiny are assumptions made concerning test standardisation, appropriate analytic techniques and meaningful interpretation of performance. The Saving Brains programme has provided a unique opportunity to push the boundaries of the debate forward.

In this study we have explored the association between the psychometric properties, reliabilities, of the tests selected for impact evaluation, and different approaches to test preparation. The data drawn from multiple linguistic and cultural settings within the Re-Enrolment programme has provided strong support for the psychometric integrity of contextualised testing protocols. In correspondence with other literature on test adaptation, the experience of the different study groups has shown that adaptation does not generally compromise, and may even enhance, the reliability of contextualised measures.

A novel aspect of the current analysis was the exploration of the contribution of distinct aspects of the adaptation process to enhancing reliability. The most cost effective changes associated with higher reliability indices were procedural. That is alterations to test administration and the scoring systems. The least effective changes stemmed from changes to content, to verbal and visual targets.

Changes to content are, however, made more frequently, possibly because the unfamiliarity of both verbal and visual items is more readily observable. The failure of children to respond, and respond correctly, to material with which they are not familiar translates into missing data, and low scores. Assessment teams commonly noted this problem, and responded with substitutions and omissions. However, the benefits of improved test-retest reliabilities must be balanced against the more limited achievement of internal consistency. These mixed observations following the process of translation and substitution may substantiate the concerns of those who advocate against test adaptation. Specifically the concern is that item selection in published tests has occurred over a lengthy and detailed process of development, to reflect a specific mix of elements of a concept, and progress through skill levels. Selecting substitutions that accurately reflect the original combination of items will understandably be difficult, especially in the time commonly available to make these changes. Arguably the favoured strategy should be to maintain the original content, as far as possible to address semantic equivalence. But when material is sufficiently unfamiliar to be unable to elicit any response at all, then maintaining item equivalence comes into question. The conclusion seems to be that, as changing content is not an easy

strategy to implement, it should be acknowledged that when it does become necessary detailed piloting and analysis is required in order to evaluate test equivalence and integrity.

Procedural changes, on the other hand, may be more easily modified to maintain fidelity to a formalised standardised approach. By making, often small, changes to the instructions given, and or the methods through which children or parents provided their responses, greater access to, and engagement with material, and consequently also to higher indices of reliability are observed. What this data suggests is that greater attention should be paid to addressing contextual differences in the social context of test administration. The data from these studies suggest that future test applications need to consider the true meaning of a standardised approach, which is to provide a standard approach to test administration that optimizes individual performance. We should be concerned more about the efficiency of test administration than about employing a regimented system that reduces understanding of the intention of the test in a new context. Whether you use chop-sticks or a knife and fork, you are, after all, still eating.. Perhaps it is time to apply the conclusions of Serpell's seminal study on this issue published in 1979.

Given the stage of analysis reached by the study teams during the preparation of this report, it was not possible to access data on test validity, and thus make an assessment of conceptual equivalence associated with different preparation strategies. Preliminary data, as well as the observation of greater difficulties in adapting and maintaining reliability measures in the executive function tasks (EF), suggests that they way children respond to tests in these new settings need more detailed investigation. It may be that the less differentiated factor structures, and the less consistent response to the EF tasks (skills and abilities that are considered highly sensitive to the effects of early brain insults), are actually reflecting an unfamiliarity of the testing situation. We need to consider both novel methods of testing, as well as novel tests.

The data has identified therefore a variety of challenges to the assumption that it is possible to deploy a common test across all settings.

Guidelines for the Future

The experience of the multiple studies surveyed highlighted important challenges to the process of test application in different settings, from which the following methodological and theoretical principles and concomitant actions are derived.

1. **The application of an assessment battery requires more than just tests**. **It also requires a testing infrastructure to support data collection, analysis and interpretation.** Clearly the language of delivery has implications for accessibility of the material, preparation time, and for the selection of a suitable assessment team. The identification, training and retention of a suitable team is central to the application of the principles of standardization.

Test preparation and delivery begins with access to appropriate assessment resources. This includes sourcing of tests with either a track record in the community, or that can be freely contextualized. Assessment also requires access to a suitable place that is conducive to making a child feel comfortable, and minimizes distractions. Those engaged in child assessment should develop a system that addresses both extended preparation time, implementation, as well as ongoing supervision to maintain quality amongst teams relatively inexperienced in assessment practices.

In constructing this system issues of sustainability and continuity can be addressed by building upon the resources developed during each testing round , as well as planning for accreditation of teams. To retain and build upon assessment skills, developing a career path and for those trained in assessment would also contribute to filling the large gap in resources for child support services lacking in many majority world settings.

2. Instrument measurement properties should be proven and not assumed. After installation all measurement instruments require re-calibration to take into account possible environmental influences such as atmospheric pressure or temperature. So too instruments used to measure human behaviour and functioning should be subjected to re-calibration to take account for language specific biases as well as culture specific cognitive styles and social customs.

In determining possible influences on 'calibration' we need to explore the influence of contextual factors such as the level of urbanization, and underlying literacy levels. Through a co-ordinated programme in data sharing the implications of these re-calibrations can be clarified, and used to develop appropriate normative tables.

3. **Comparison between sites should focus on statistical effects, rather than raw or standardised scores.** At the content level there are obvious limitations to the ability to apply a universal tool. This should not be seen as a limitation drawing conclusions across studies and settings. An index is merely a means of measuring constructs, the means, rather than the end in itself. An appropriate index need only provide an accurate method of measuring of concepts of interest, the Core Metrics. Other methodologies are available to summarise the effects they illustrate.

Data from multiple indices can be summarised at every level of specificity, depending upon the communality of the underlying variance. Statistical methods that enable comparison, both within and between populations, provide a stronger underlying logic to summated data, and may also better address the limited inherent validity of single measurement scales alluded to earlier. Future initiatives need to address the way in which data is collected, to enable a common analytic methodology.

Effect sizes, for example, are now a requirement by the American Psychological Association for publication in their journals. There are also several articles that refer to this approach as way to carry out meta analyses (*see link provided in References).* If a common methodology based upon effect size is calculated, then data across sites and studies can be submitted to a meta-analytic approach that overcomes the necessity to look for that single gold standard test.

4. **Access to and Extension of Materials that measure Human Capital Formation in the Majority World.** Through wider application and usage of tests, and a sharing of that experience over time, we can develop a repository of well-translated, 'calibrated' tests. This task should be approached not only with scientific rigour, but also with a significant degree of innovation and creativity.

Currently published tests are largely based upon Western philosophical definitions of intelligence. While recent modifications and additions to the western test library have recognised broader definitions, and multiple intelligences, they do not incorporate non-western constructs in the underlying framework of test design. Thus there is little known of the ability of assessments to monitor key features of adaptive or independent functioning that necessarily predict development skills important for functioning within different socio-cultural-economic-linguistic contexts.

The identification and definition of appropriate Core Metrics will need to take into account issues of contextualisation in terms of accounting for life stages, as they have done in the changes made between the RE and SI programme rounds, for cultural niches, as well as to account for the specific questions raised by the focus of an intervention.

In expanding psychological assessment in the majority world we need to not simply rely upon a limited western model. To address the demands of different environments there is room for innovation, adjustment, change, and improvement to the current Core Metrics Framework

"The SB community can act as a "community of practice," equipped with a well-tailored, integrated kit of tools and capacities designed for collective impact in human development (e.g. appropriate metrics, analytical tools, relevant adaptation skills, theory of change). These tools and skills, which might develop out of focused working groups, need to be made accessible to the whole community. Their use and value will then be demonstrable through shared learning, and a commitment to the collective enterprise." JR 2015

References

Abubakar, A., van de Vijver, A.J.R., van Baar, A., Mbonani, L., Kalu, R., Newton, C.J.R. & Holding, P. (2008) Socioeconomic Status, Anthropometric Status, and Psychomotor Development of Kenyan Children from a Resource-Limited Setting: A Path-Analytic Study. Early Human Development 84 (9) 613-621 (4)

Alcock K, Holding P, Mung'ala-Odera V, Newton CRJC (2008) Constructing tests of cognitive abilities for schooled and unschooled children Journal of Cross Cultural Psychology 39.529-551 (2)

Chione, D.P.S.B. & Buggie, S., E., (1993) Memory performance of African oral historians. *Journal of Psychology in Africa.,* 1(5): p. 123-135.
Cohen, J., (1992) A power primer.*Psychological Bulletin,* 112(1): p. 155-159.

Kearins, J. (1986), Visual spatial memory in aboriginal and white Australian children. Australian Jnl of Psychology, 38: 203–214.
doi: 10.1080/00049538608259009

Serpell, R., (1979) How specific are perceptual skills? A cross-cultural study of pattern reproduction.*British Journal of Psychology,* 70: p. 365-380.

Vierhaus M, Lohaus A, Kolling T, Teubert M, Keller H, Fassbender I, Freitag C, Goertz C, Graf F, Lamm B, Spangler SM, et al. (2011) The development of 3- to 9-month-old infants in two cultural contexts: Bayley longitudinal results for Cameroonian and German infants. European Journal of Developmental Psychology 8(3): 349–366.10.1080/17405629.2010.505392

Wagner, D.A., (1974) The development of short-term and incidental memory: A cross-cultural study.*Child Development,* 45: p. 389-396.

General References on Effect sizes: http://ies.ed.gov/funding/pdf/effectssize.pdf

APPENDIX 1 – TESTS SELECTED RE STUDIES

Core Metrics	Construct (s)
Estimate of General Intelligence	
WASI-2 (2)	Verbal and Performance IQ
WPPSI (2)	Fluid reasoning, Receptive & Expressive language
WISC IV	Verbal comprehension, perceptual reasoning, working memory, processing speed
Information	Verbal Ability
Block Design	Visuo-spatial Processing
KABC II	Non-verbal Intelligence
Atlantis	Learning
Hand Movements	Memory/Sequential Processing
Footsteps	Simultaneous Processing
Story Completion	Planning
Sangian	Learning, Sequential Processing, & Short-Term Memory, Visuo-Spatial Processing,
Executive Function (Batteries)	
Sangian	Working Memory, Attention, Inhibition, Cognitive Flexibility
Executive Function Battery	Memory and Cognitive Flexibility
Executive Functioning	Inhibitory Control, Working Memory, Cognitive Flexibility
Executive Function (Attention)	
Stroop Numbers	Inhibition
Digit Span	Auditory Attention and Working Memory
Visual Search Dual Task	Sustained Attention
Test for Attentional Performance	Multiple Dimensions of Attention
Day and Night Test	Inhibition, Working Memory
No-Go	Inhibition
Shift	Attentional Shift
People Search	Sustained Attention

Core Metrics	Construct (s)
Executive Function (Cognitive Flexibility /Working Memory)	
Rey Osterrieth Complex Figure-Recall	Visual Working Memory
Rey Osterrieth Complex Figure-Copy	Visuo-Motor Organisation/Planning
Spatial working memory	Working Memory
Digit Span	Auditory Attention and Working Memory
Dimensional Change Card Sort	Cognitive Flexibility
Presence of Behaviour Problems	
Strength and Difficulties Questionnaire	Internalising and Externalising Behaviours
Child Behaviour Checklist	Anxiety/Depression, Rule-Breaking, Aggressive Behavior
Emotional Stroop	Dominant/Submissive
Spence Children's Anxiety Scale	
Child Behaviour Profile	Emotional Stability, Conduct problems, hyperactivity/Inattention, pro-social behaviour
Socio-Emotional Assessment	Socio-Emotional and Behavior
Conners 3	ADHD
Adult Behavior Check List/ Adult Self Report	
Literacy	
Literacy	Various Sources
Numeracy	Various Sources
ACER - Applied Reading (2)	Website
Literacy	Described in Published Article
BSRA-3-Letter Recognition	Published Test
Others	
Kilifi Naming Test	Expressive Vocabulary
Shivgarh Motor Proficiency Test	Fine and Gross motor function
Visuo Motor Integration	Visual-Motor integration
Living Habits of Young Adults	Health Habits
Kidscreen 52	Quality of Life
Inventory Of Parent and Peer Attachment	Social Networks