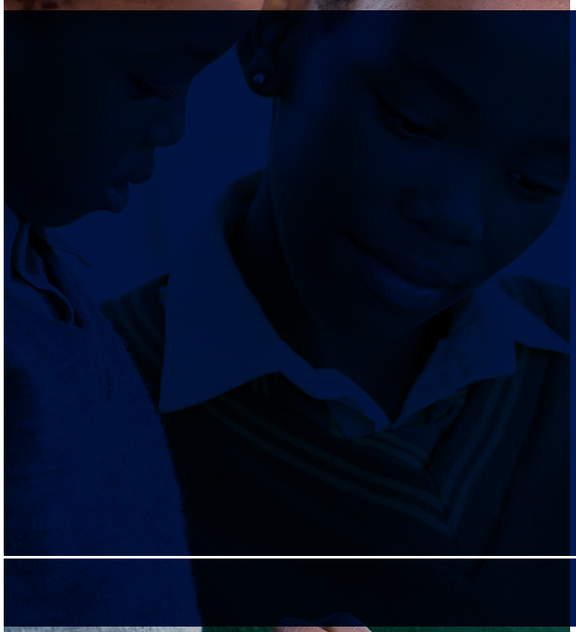

Contextual issues in the assessment of children's learning

Briefing Note

Findings from a Rigorous Literature Review of Assessment of Literacy and Foundation Learning in Developing Countries.

This policy brief discusses some of the emerging contextual issues in literacy and language assessment, which were raised in a DFID-funded rigorous review of children's foundation learning in developing countries: Nag (2017) ***Assessment of Literacy and Foundation Learning in Developing Countries*****.**

The need for at-scale measures of learning that are contextually appropriate

In the past two decades, there has been a huge expansion in the range of measurement tools available to assess foundation learning and literacy in developing countries. The vast majority of the measures are researcher-developed tools, which are generally used in smaller samples, and in response to a specific research aim. However, there is an increasing preference for learning assessments that can be applied at-scale, with a view to address local educational policy and programming issues. In both cases, tests with demonstrated good psychometric properties in Western contexts, are increasingly being adapted for use in developing countries.

This raises concerns about the appropriateness of an assessment when transferred from one context to another, particularly if the test is to be used for population-level or programme monitoring. Most tests cannot be feasibly scaled up, as they are expensive, take a long time to administer and require substantial technical skill. The findings of the literature review further suggest that the purpose of learning measurement at-scale is often not clearly defined.

Where at-scale assessments can add value depends on the extent to which they are able to inform questions regarding the quality of education. Nag (2017) suggested that developing meaningful measures to inform educational policy decisions is currently a challenge. For example, the review showed that the widely used print knowledge tasks, which measure aspects of emergent literacy, often do not capture individual differences in children's orientation to print in settings where poverty levels are high. This is because children living in poverty may have limited exposure to print materials and reduced access to instruction. The children who are tested often end up with the same low or zero scores, which does not produce the level of test score variance needed to differentiate between the children.

Such contextual errors of measurement can be avoided by taking into account the local setting in which a test is to be administered. Alternate materials that make sense in the local context (e.g. the use of stimuli that draw on local print products to test print knowledge), may improve the relevance of measures, and in turn, the usefulness of the test data.

The need for at-scale measures of learning that are contextually appropriate

Language characteristics and cultural factors significantly affect the performance of learning assessments. To ensure that an assessment provides valid and reliable information, it needs to be designed for, or adapted to, the linguistic and cultural context in which it is being administered. This is a particular challenge in multi-cultural/lingual testing contexts, or contexts where children are learning in a language other than their home language.

The results of the review showed that there is significantly better reporting of localisation among researcher-developed tools, but that commercial test adaptation processes are poorly documented. A large proportion of the studies using adapted versions mention translation and the use of consultations with local experts to select and translate items, but do not detail the actual translation procedures.

This is a major concern as test localisation is not only a matter of literal translation of the test, it involves ensuring that the test format, its content and the testing process are familiar and meaningful for test-takers. Piloting is a way of ensuring that test design decisions do not contribute to costly errors of measurement. However, for a large proportion of reviewed studies, pilots were either not carried out or not reported. The consequence of such oversight, again, relates to the usefulness of the resulting data. For example, the results of the review showed that in languages like English and Spanish letter naming tasks generally discriminate poorly between children, as most children master this skill by Grade 2. Unless an

appropriate combination of single letters (e.g. "t") and more complex letter combinations (e.g. "-tion") are included, this task may not provide useful information. Related to this, letter sound tasks, were shown to discriminate poorly as well, because this skill is often ignored in formal instruction, and thus few children are able to answer these questions correctly. Without sufficient variation in the sample, the resulting test data cannot be used for analysis of individual differences and, in turn, cannot fully inform policy and programme decisions.

Improved transparency in the reporting of the robustness of learning measures

Overall, documentation of the statistical reliability and validity of learning assessments is found to be very poor. In general, internal consistency reliability coefficients are reported, but only as a way of establishing overall test performance and not in conjunction with any other substantive information to argue the robustness of the measure used. In some cases, even reliability estimates are excluded from reporting. For example, reliability information is provided for only 30% of the measures used to assess reading accuracy.

The review finds that psychometric properties of the test items are rarely analysed. Such analyses would be needed to confirm possible sources of measurement error. Issues that can be addressed using psychometric techniques include: whether the items on the test adequately target the range of children's proficiency levels; whether some items do not capture individual differences in proficiency; the extent to which each item effectively contributes to what is being measured; identifying sources of bias that may be present in the test data, cultural, linguistic or other contextual factors (geographic, demographic, socioeconomic); and whether there are redundancies or dependencies between questions on the assessment that need to be resolved to improve the accuracy of reliability estimates. Adding, removing or modifying questions in the test to address such contextual sources of error, through piloting and statistical analysis, will improve the accuracy of the measures, and in turn the validity of test score interpretations.

Procedures for making qualitative judgements on robustness are also underreported. The review found that procedures for establishing measurement equivalence do not seem to be informed by any guidelines or standards. Establishing the robustness of learning measures should not be limited to statistical analyses of the test data, but may also require qualitative information, such as: how the construct being measured is understood across the contexts of study, whether the questions on the test share a common meaning across the study contexts, and the extent to which administration procedures, the test format and stimuli used during testing are familiar to test takers. As elegantly summarised in the review on the difficulties of establishing equivalence in picture stimuli between children of different socioeconomic and cultural backgrounds:

> *"…pictures may be ambiguous and the reason for this could be several: pictorial representations may be alien for children with exceptionally low print experience, the pictorial idiom may be difficult for children to understand or the visualisation may be outside their lived experience."*

Main messages

There has been a steady expansion in the range of measures used to assess foundation skills in developing countries, with a growing preference for tests that can be used at-scale. Deciding what knowledge to assess at-scale should be based on what information will shed light on the quality of education, and this requires an understanding of the contexts in which assessments take place.

Developing countries present diverse cultural and linguistic contexts, and there is a need to ensure that learning assessments are designed to suit each context. However, the process of test localisation remains largely unreported in the literature. This is a particular concern when tests are adapted from one cultural/ linguistic context to another.

The vast majority of studies report on children's learning with no supporting qualitative or statistical evidence of the robustness of the measures used to assess learning across different contexts. The few discussions about the psychometric properties of tests are generally limited to indicators of test reliability.

This brief was drafted by Nardos Tesfay, based on the Rigorous Literature Review of Assessment of Literacy and Foundation Learning in Developing Countries, by Sonali Nag.

About Oxford Policy Management

Oxford Policy Management (OPM) is one of the world's leading international policy development and management consultancies. We enable strategic decision-makers in the public and private sectors to identify and implement sustainable solutions for reducing social and economic disadvantage in low- and middle-income countries supported by offices in the UK, Bangladesh, India, Indonesia, Nepal, Pakistan and South Africa.

For further information, visit **www.opml.co.uk**