Development of Assessments of Reading Ability and Classroom Behavior

A report prepared for the World Bank

by

Matthew Jukes,

Shaher Banu Vagh

And

Young-Suk Kim

Harvard Graduate School of Education,
Appian Way, Cambridge MA 02138

Email: matthew_jukes@gse.harvard.edu

Final Draft – 15th September 2006

Introduction

The World Bank seeks to develop a standardized approach to assessing educational outcomes and processes in order to generate comparable data across countries conducting impact evaluations of innovative policy and program interventions at the primary school level. . Initially, two assessment tools will be developed. The first is a test of Grade 2 reading ability. The second is a protocol for the observation of classroom behavior and activities.

The design of assessment tools will ensure reasonable robustness to cultural and national differences in the target measures. For the reading test, different skills are requiring for competence in different languages, and curricula vary in the demands placed on pupils. Thus reading tests will focus on commonalities across languages and across curricula. Classroom behavior also varies with context due to differences in planned classroom activities and to cultural differences in behavior of teachers and their pupils. The observation schedule will seek to produce valid assessments across these various contexts.

The report contains two sections. The first section reviews approaches to the assessment of reading ability and reports the results of a pilot study in Kenya to develop a standardized approach to the assessment of reading ability. The second section reviews approaches to assessment of classroom behavior and described a new method currently being piloted in Kenya.

Review of Reading Assessments

Introduction

The acquisition of reading is a developmental and multifaceted process. Good readers, no matter what language they read, demonstrate the ability to (a) identify the words on a page and (b) make meaning of the text being read. Consequently, many reading measures assess comprehension directly. However, the meaning-making process - i.e. reading comprehension – is dependent upon and indicative of the mastery of a number of underlying and related skills, such as, knowledge of the units of a writing system, knowledge of the ways written units map on to sound units, and vocabulary knowledge, background knowledge, and grammatical knowledge. For successful reading comprehension, words in connected text need to be decoded with speed and accuracy such that limited cognitive resources, such as attention and memory can be allocated to text comprehension. Hence, alternatively many reading measures assess the constituent processes of reading. This latter approach has advantage for cross-linguistic reliability and for the assessment of early reading skills (discussed below).

Reading research over the past several decades has compellingly demonstrated the importance of the constituent early literacy skills - such as letter knowledge, letter(s)-sound mappings, and vocabulary knowledge - to later academic achievement. Evidence from several countries indicates that children who are poor readers early on are more likely to struggle with reading and as a consequence with all of their academic subjects for the rest of their schooling years (Clay, 1991; Cunningham & Stanovich, 1997; Ferreiro & Teberosky, 1982; Juel, 1988). This is a finding of great concern given the low academic achievement levels reported for children in early grades in poor, developing societies. This underscores the importance of early literacy instruction and indicates that there is an overwhelming need to monitor early literacy development and the attainment of early reading goals via appropriate reading assessments in order to effectively inform the design of standard and/or intervention programs.

Theoretical Context

Reading comprehension has received wide research interest, and has proven to be a contentious topic particularly in the Western, developed hemisphere and notably with the English language. Currently, there is converging theoretical argument and empirical evidence on the role of "fluency" as a necessary condition as well as outcome of reading competence in English (see special issue on fluency of the Scientific Studies of Reading, vol 5, 2001). This is based on the premise that comprehension, the goal of all reading, relies on various skills and knowledge, basic among which are the "accurate" and "efficient" decoding of words. The automaticity of lower-level skills ensures that limited cognitive resources, such as attention and memory can be freed and allocated to the higher-level skills of meaning-making (LaBerge & Samuels, 1974; Perfetti, 1977, 1985). In the process of defining "fluency" researchers have implicated a host of multi-dimensional processes and componential skills that need to be well-orchestrated for successful reading (Wolf & Katzir-Cohen, 2001). These involve the building of reading fluency within and between several processes – phonological, orthographic, morphological and semantic processes (see Box 1) that operate at different levels – letters, letter combinations, words, and connected text. Hence, Wolf and Katzir-Cohen conclude "that reading fluency involves every process and subskill involved in reading" (p.220).

Box 1: Some definitions

Phonology refers to the ways the sounds of a language function

Orthography refers to the ways spelling patterns are represented by a writing system

Morphology refers to the way words are formed (inflections, derivatives, compounds) and are related to each other

Semantics refers to the ways in which language conveys meaning

Much of the research that expounds the importance of fluency – i. e. speed and accuracy of decoding words in *connected text* to reading comprehension - has been focused on the English language which has an exceptionally irregular orthography. While there is a growing body of research examining the processes underlying reading acquisition in various other orthographies, the bulk of this research has focused on processes that determine the successful acquisition of word decoding skills in isolation and less on the decoding of words in connected text and its association to reading comprehension (Katzir, Shaul, Breznitz, & Wolf, 2004; Saiegh-Haddad, 2005). Nevertheless, this focus on word recognition is of critical importance as reading comprehension begins at the word level and "unless the processes involved in individual-word-recognition operate properly, nothing else in the system can either" (Adams, 2004, pg. 1219).

Several studies of children diagnosed with dyslexia and learning to read in regular orthographies have however been illuminating. This work suggests that due to the regular grapheme-phoneme correspondences of regular orthographies, like Italian, children with dyslexia can attain a high level of accuracy in reading. However, the core of the problem centers around speed, i.e. slowness in reading (Zoccolotti et al, 1993 & Tressoldi et al., 2001).

The next few sections present an overview of some of the key components or the reading subskills that have been the focus of much research for studying and assessing the acquisition of reading and the correlates of oral reading fluency. However, first a discussion about the various writing systems and the challenges they present for young learners.

Languages and Writing Systems:

Writing systems map graphic units to language units such that spoken words for the given language may be represented in writing. The elementary units of writing can correspond to phonemes as is the case for alphabetic writing systems like English; or to syllables as is the case for syllabaries like Japanese Kana; or to morphemes (smallest units of meaning) as is the case of logographic writing systems like Chinese or the

Japanese Kanji. Moreover, the graphic units of a writing system can map on to more than one phonological unit as is the case for alphasyllabaries like Hindi where graphic units correspond to phonemes or syllables. Hence, beginning readers are likely to be faced with different challenges in different languages, such as the number of graphic units that need to be learned, for example, English has 26 graphic units, while Hindi has about 52.

Moreover, alphabetic writing systems differ along a continuum of deep to shallow reflecting the consistency with which letters map on to sounds. For shallow orthographies like Swahili, Italian, Finnish, the mapping is highly consistent: that is, each writing unit or a combination of them is most likely to represent a single sound unit and vice versa. On the other hand, for deep orthographies like English, the mappings are highly variable: that is, one writing unit may represent two or more sound units depending on its position in a word and conversely, the same sound unit can be represented by two or more different writing units.

Knowing the letters:

Reading to learn in all languages begins with the learning of the elementary units of writing. For alphabetic writing systems these are 'letters', for syllabaries these are 'syllables', and for logographic scripts these are 'morphemes'. This is because these elementary units of writing – letters, syllables, morphemes are the most accessible unit in print. This is also the reason that this is one of the components that lends itself easily to assessment.

Much correlational evidence for the English language indicates that a child needs to know the alphabet in order to learn to read as this initiates the first level of reading acquisition by providing a basis for analyzing words into their constituent parts, and in so doing provides a link to phonology. For English, knowing letter names and or sounds have been identified as a skill strongly positively correlated to early reading (Chall, 1996; Clay, 1967; Ehri, 1979; Scarborough, 1998; Sénéchal & LeFevre, 2001; Share, Jorm, Maclean, & Matthews, 1984) . Scarborough's (1998) review suggests that letter knowledge accounts for one third of the variance in reading performance from first through third grade. In addition to the knowledge of letter names or their sounds, the

speed in identifying letters has also been identified as an important skill for its close association with school reading achievement (Hecht, Burgess, Torgesen, Wagner, & Rashotte, 2000), and as it indexes the efficiency of phonological and orthographic processing and the temporal coordination between them (Wolf, Bowers, & Biddle, 2000).

For children learning to read regular orthographies, unlike English, the names of letters correspond to their sounds and for the most part are associated with a single sound, thus presenting young children with a relatively simpler task in gaining mastery of letter-sound correspondences. Moreover, letter knowledge in regular orthographies like Finnish and German directly precedes skills related to phonological tasks such as the assembly of letter sounds to form letter combinations and words (Holopainen, Ahonen, & Lyytinen, 2001; Näslund & Schneider, 1996). For regular orthographies too, such as Arabic, Dutch, Latvian, and Turkish letter knowledge is as an important correlate of reading skills (Oney & Durgunoglu, 1997; Saiegh-Haddad, 2005; Sprugevica & Hoien, 2003; Wagner, 1993; Wesseling & Reitsma, 2000)

Assessing Letter Knowledge:

The assessment of letter knowledge is one of the most widely applied, useful, and easy to administer assessments. Letter knowledge has been assessed on the basis of children's ability to (a) accurately provide letter *names*, (b) accurately provide letter *sounds*, for irregular orthographies like English, and (c) fluently name the letters. All of these skills have been identified as important correlates of reading ability (Foulin, 2005). Given the emphasis placed on automaticity of lower level skills for fluent reading of words in connected text, we recommend the use of a letter reading fluency assessment. For this, children are asked to read aloud as quickly as they can the names or sounds of the letters presented in random order. The number of letters correctly identified in the span of 60 seconds provides a score of *letter reading* fluency.

It is of interest to note that in a multi-country, cross-language comparative study, Seymour et al (2003) did not find any systematic variations in children's letter knowledge in both accuracy and speed as a function of orthographic depth, and syllabic complexity, factors that were associated with differences in children's word decoding abilities (see

discussion below). This further corroborates that irrespective of orthographic depth, letters are the most salient units in a writing system.

However, when comparing children's letter reading fluency across languages, for example Swahili and Hindi, it should be noted that these two languages differ in the total number of elementary writing units, while the Swahili alphabet comprises 24 basic units, the Hindi alphasyllabary comprises about 52 basic units.

Knowing the mappings between written units and sounds:

Three decades of reading research has provided ample and robust evidence for the critical role of phonological awareness in the acquisition of learning to read in English and several other languages (Goswami & Bryant, 1990). Phonological awareness refers to the understanding that words can be segmented into its constituent units, such as syllables and phonemes. Phonemes represent graphemes and knowledge of grapheme-phoneme correspondence rules (phonological recoding) hence translates into learning to read in an alphabetic writing system.

Languages based on alphabetic writing systems vary not only across a continuum of shallow to deep, depending on the consistency with which the units of writing map on to units of sound, but also differ in the complexity of their syllable structures. The learning of the mapping of grapheme-phoneme correspondences is relatively simpler for regular orthographies with simple syllable structures like Italian than it is for regular orthographies with complex syllable structures like German. However, the learning of grapheme-phoneme correspondences for irregular orthographies that also have a complex syllable structure, like English, is the most challenging.

Comparative studies have shown that the rate of acquisition of grapheme-phoneme correspondences varies as result of orthographic depth (Seymour, Aro, & Erskine, 2003). Essentially, orthographies, that are either inconsistent in reading like Danish, or in spelling like Hebrew, and French or in both reading and spelling, like English, present greater challenges in mastering the code than shallow, regular orthographies. For example, young children learning to read in English take about two-and-a-half times longer to attain basic foundational skills compared to young children

learning to read other orthographically shallow orthographies like Finnish, Greek, Italian, Spanish, or German (Aro & Wimmer, 2003; Seymour, Aro, & Erskine, 2003). Furthermore, even among the group of orthographies that are considered to be regular, the acquisition of basic reading skills varies as a function of the relative depth of the individual orthographies.

Assessing phonological processing skills:

The ability to manipulate and distinguish the sounds of a language - i.e. phonological awareness - have been widely assessed in younger children by tasks that require children to decide whether two words rhyme or if they start or end with the same sounds, to delete initial or final sounds, and count the sounds in tapping tasks.

The ability to apply the grapheme-phoneme correspondence rules - i.e. phonological recoding - is commonly assessed through *nonword* reading. This is suitable for slightly older, primary grade children – middle or end of first-grade onwards. The reading of nonwords requires children to apply the rules of letter(s)-sound correspondences as other strategies such as lexical access, or the holistic recognition of words cannot be successfully applied since the words are not real, familiar words.

Standardized assessments as well as researcher developed assessments for the measurement of both phonological awareness and phonological recoding abound in the literature. For the most part these assessments are time consuming, require trained examiners to administer, and are harder to gain reliability in the coding of errors. An additional challenge for cross-linguistic comparisons is the development of comparable items across languages. All of this is further compounded by the differences across languages in their syllabic complexity, i.e. simpler open syllables, CV that prevail in Italian, Spanish, etc. or more complex closed syllables, CVC as well as even more complex consonant clusters, like CCVC that prevail for German, English, etc. Although indispensable to basic research to help illuminate the underlying processes of reading, these assessments are not recommended for cross-linguistic, cross-country program evaluations due to the challenges of test development, training, and administration as noted above.

Oral Reading Fluency:

Oral reading fluency indexes the *accuracy*, *speed*, and *prosody*[1] or expression with which written text is read. Several studies in English have provided evidence for oral reading fluency as a robust index of reading comprehension and one that reliably differentiates good readers from poor readers (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Good, Simmons, & Kame'enui, 2001; Stanovich, 1991). This is because efficient and accurate reading of words in connected text signals the efficient and accurate functioning of lower level foundational skills such as knowledge of the writing system at the level of letters, letter combinations, and words. This includes the ability to differentiate and manipulate the sounds of a language (phonological awareness), and the ability to map print units on to speech units (phonological recoding). However, reading words in connected text taps more than just the ability to decode words efficiently and accurately. If the effective decoding of words were all that mattered for reading comprehension then reading words in list form would serve as an adequate measure for determining children's reading ability levels. However, as has been demonstrated by a growing body of research, reading connected text brings into play several subcomponents at both *lower* and *higher* levels of processing. Hence, in addition to efficient word decoding – a lower level of processing – it also involves higher level processes, such as the efficient integration of propositions so as to build a mental representation of the text's message. It is for this reason that oral reading fluency of words in context is a far superior measure of reading comprehension than word list reading {Jenkins, 2003 #46; Jenkins, 2003 #65}, at least during the primary grades.

Working memory, which is a system of limited capacity, plays an important role in storing and manipulating information derived from lower and higher level processes associated with reading and reading comprehension. In order to decode, children must be able to retain the phonological representation of orthographic units in working memory until phonological assembly and lexical access has been achieved. Moreover, in order to

[1] Most oral reading fluency measures focus on two of these three components of fluency, namely 'accuracy', and 'speed'. This is because the assessment of 'prosody' in a reliable manner presents many challenges.

comprehend, children must be able to remember the words in the sentence that has just been read, retrieve relevant information from preceding sentences, and integrate information derived from current and preceding text. Hence, given the host of processes operationalized by working memory relative to its limited capacity, measures of memory form a good index to differentiate good readers from poor readers (Abu-Rabia & Siegel, 2002; Geva & Siegel, 2000; Just & Carpenter, 1992). It has been argued that differences in working memory capacity are caused by phonological processing difficulties - the ability to hold and manipulate phonemes in memory (Crain & Shankweiler, 1988).

Assessing oral reading fluency:

A simple yet useful method of assessing oral reading fluency is by asking children to read aloud from curriculum-relevant texts and counting the number of words accurately read within a span of 60 seconds. The number of words read correctly per minute indexes the child's oral reading fluency score. For the purpose of cross-linguistic comparison an alternate and preferable approach, although not one without some limitation is to use the *syllable* as the unit of analysis. In other words, children's oral reading fluency scores can be estimated as the number of accurately read syllables within a span of 60 seconds.

Two types of standardized oral reading fluency assessments are widely reported in the literature: (1) commercially developed assessments, and (2) curriculum-based assessments. Most commercially developed assessments have the advantage that they provide norms that help understand children's reading ability relative to their peers. Additionally, normative data represent performance standards so as to understand the expectation for a particular age and grade level. Curriculum-based assessments, on the other hand are drawn directly from children's prescribed curriculum and therefore provide greater instructional and curricular validity and have also been found to have favorable psychometric properties (Deno, 1985; Fuchs & Fuchs, 1999; Fuchs, Fuchs, & Maxwell, 1988) for screening individual students as well as monitoring their progress. There are two main advantages of curriculum-based assessments: (a) many alternate test versions are easily available, and (b) they are inexpensive to produce. Moreover, it is possible to obtain normative data based on curriculum-based assessments as reported by

Hasbrouck and Tindal (2006). An additional important advantage of such assessments for cross-country and cross-linguistic comparisons is that children learning to read in different languages and in different contexts will have the opportunity to read texts that were originally written in their languages, are part of their prescribed curricula, and presumably represent familiar cultural knowledge rather than be subjected to read translations from other languages that may introduce construct irrelevant linguistic and cultural factors (Bonnet et al., 2001).

Although, oral reading fluency assessments require more time than a group administered assessment, relative to other individual assessment it is quick and simple to administer, taking only a few minutes.. However, it does require examiners to be trained so that they can reliably identify children's reading errors. Where classification of errors is not an aim of the assessment, it is not necessary to train examiners to classify errors into their sub-types (errors of omission, substitutions, etc). However training in the different types of errors will not only help alert examiners to reading errors they are likely to encounter but also to code these in a consistent and reliable manner.

The feasibility of monitoring children's reading development on a large scale by the administration of individual tests has been demonstrated by the assessment work carried out by several organizations. Notable among them is the work of the Indian non-governmental organization, Pratham, which has recently completed a massive nationwide program in India to assess children's reading ability in the age group of seven to fourteen attending all government and private institutions in rural India (Pratham, 2005). Additionally, Pratham has also been conducting large scale assessments of all children attending its various educational programs in urban centers using assessments that test children's knowledge of the writing units, ability to decode words in list form and ability to read short passages. However, as traditionally has been the focus, especially of educators and teachers, the focus is primarily on *accuracy* with rough subjective guidelines for speed. Moreover, the focus is only on decoding ability and not on comprehension. Coding of reading ability is also somewhat subjective – reliability and validity have not yet been demonstrated. For these reasons, the Pratham reading tests are not suitable for reliable cross country assessments in their current form. These reading

assessments, which are similar in structure to the assessments proposed in this paper can be accessed at http://www.pratham.org/aser-report/Page%209.pdf.

Cross-linguistic issues in oral reading fluency (ORF)

There is both theoretical and empirical support for the use of ORF as an index of reading comprehension. However, as noted above, much of the evidence is based on studies in English. This presents the problem of generalizability as English has an irregular orthography relative to other orthographies like Italian, Spanish, or Swahili. The variation in grapheme-phoneme correspondences for English implies that some word knowledge is necessary for reading fluency as not all words can be decoded via phonological assembly. For instance, a reader has to know that 'ough' is pronounced differently in 'tough' and 'though'. This dependence on word knowledge suggests that reading fluency is to some extent indicative of word knowledge that facilitates comprehension. However, for regular orthographies the grapheme-phoneme correspondence rules are more consistent as a result of which novel words can be decoded directly via the phonological route not necessarily involving comprehension. As a result, for regular orthographies, oral reading fluency many not be as highly correlated with reading comprehension as it has been demonstrated for English. The lack of evidence on this issue is addressed with a pilot study in Swahili, a regular orthography, that examines the association of oral reading fluency to reading comprehension measures and other validated measures of reading ability in Kenya. These results are discussed in Appendix A.

Outcome measures

Oral fluency is often reported as the number of words read in a fixed period of time. Using words as the unit of measurement may present problems for cross-linguistic comparisons. The length and complexity of common words varies from one language to another and thus the rate at which they are read and pronounced may vary accordingly. An alternate strategy would be to report outcome measures in smaller linguistic units.

There is less cross-linguistic variability in syllable length than in word length, and less still in the length of phonemes. Table 1 reports the number of words, syllables and phonemes found in reading passages from Grade 1 text books in Spanish from Peru {Abadzi, 2005 #15} and in English and Swahili from Kenya (study reported in Appendix A). Table 1 shows that, as expected, the cross-linguistic variability in number of linguistic units found in a typical Grade 1 reading passage declines with the size of the linguistic unit. There is almost three times more variability in number of words compared with number of phonemes

Table 1. Number of linguistic units in typical passages from Grade 1 reading textbook of three languages.

	Words	**Syllables**	**Phonemes**	**Sentences**	**Syllables/Word**	**Phonemes/Syllable**
English B	61	90	220	10	1.5	2.4
English C	63	85	221	8	1.3	2.6
Swahili D	41	100	192	9	2.4	1.9
Swahili E	41	130	240	10	3.2	1.8
Spanish	63	99	219	9	1.6	2.2
Variability*	**21.8%**	**17.3%**	**7.8%**			

*Variability is standard deviation expressed as a proportion of the mean.

Readability Formulas

For cross-linguistic and cross-country comparisons, an important consideration is the difficulty levels of the texts used in different languages and/or countries, as children's performance is likely to vary as a function of the text's difficulty level. Readability formulas are widely accepted as well as challenged by reading researchers and practitioners and at best they are regarded as imprecise yet useful tools. However, precision is a greater concern when readability formulas are used to help teachers identify readable texts for individual students than it is to make judgments about group level

performance. Readability formulas as widely used in the U.S. context have established text difficulty on the basis of syntactic and semantic complexity.

Semantic complexity is determined either on the basis of word familiarity if a frequency list of words is available or on the basis of number of syllables per word. *Syntactic* complexity is determined on the basis of number of words in each sentence of the text and estimating an average score for the text.

Given that for many of the languages in which children in developing countries learn to read and write there are no available lists of high frequency and low frequency words, semantic complexity will then have to be determined on the basis of average number of syllables per word.

A complimentary approach that can be applied along with estimating the semantic and syntactic complexity of a text is to estimate its type-token ratio. This provides an index of the *lexical diversity* of the texts. This is easily done by obtaining a ratio of the total number of unique words in the text (types) to the total number of words in the text (tokens). For example, Sentence A below has 14 words in total and 10 unique words. Therefore, its type-token ratio is 0.71. Sentence B also has 14 words in total but has 12 unique words. Hence, the type-token ratio for Sentence B is 0.86, indicating that relative to Sentence A, Sentence B has a higher type-token ratio indexing greater lexical diversity. A consideration of lexical diversity is important in the measurement of fluency as children who read passages with many repetitive words are bound to have an easier time and go through them faster, than children who read passages that are more lexically diverse as they will have to decode a greater number of different words through the passage.

A: The boy and his dog ran up to the other boy and his mother.
B: The boy and his dog ran up to the other girl and her mother.

Norms

There is absence of research that provides norms for children learning to read in developing countries. This is a hindrance given the robust finding that children vary in the rates of learning to decode words based on the characteristics of the orthography as discussed earlier. In other words, second graders learning to read in an irregular orthography perhaps may demonstrate lower oral reading fluency scores than children learning to read a relatively more regular orthography. Hence, in the absence of normative information, it is important to (a) take into account characteristics of the orthography that children are learning to read, (b) evaluate the results in keeping with the demands of the curricula, and (c) capitalize on any opportunities to collect information on general reading trends for a given language in a given country.

Oral Reading Fluency Norms:

Listed below are oral reading fluency norms available for a few languages.

English:

Norms for children learning to read English in the U.S. indicate that on average, by the end of first grade children read at a rate of 53 words per minute, by the end of second grade children read at a rate of 89 words per minute, and by the end of third grade they are reading at a rate of 107 words per minute (Hasbrouck & Tindal, 2006).

Italian:

Norms updated in 2004 for children learning to read Italian indicate that on average, second graders read at a rate of 137 syllables per minute or 65 words per minute, and third graders read at a rate of 181 syllables per minute or 86 words per minute (Tressoldi, , personal communication).

Spanish:

Norms for 2006 reported for Chilean children learning to read Spanish indicate that children classified at the high average level in first grade read 38 words per minute, in second grade read 64 words per minute, and in third grade read 88 words per minute.

Question-Answer Comprehension Tests:

Due to the dearth of basic research examining the association of oral reading fluency with reading comprehension in languages other than English, coupled with the lack of normative data, an informative measurement strategy involves a question-answer test. The questions are based on the passage that children are asked to read aloud and are orally presented. The association of oral reading fluency and comprehension can then be directly evaluated. Such a test has several advantages, chief among them is that it represents a familiar task for most school going children, as such an assessment forms a part of their routine classroom evaluations used by teachers to directly evaluate reading comprehension. In addition to informal classroom evaluation, this form of assessment is also commonly incorporated in many standardized tests. However, as noted by Fuchs et al (1988), there have also been several criticisms of such tests, chief among them are that it assesses comprehension based on selected portions of the text, and that the quality of the questions matters. For if responses can be inferred directly from the questions themselves then the test looses its credibility as a test of reading comprehension. The first criticism though is less critical than the latter for the current purpose of program evaluation. First, care should be taken in selecting passages so that they are not based on a topic that draws very specifically on world knowledge as then children familiar with the topic can draw upon their knowledge of the topic to respond to questions irrespective of reading ability. Second, care should be taken in designing test questions so that they tap information that can be provided only if the child was able to read the text rather than being inferable from the question themselves. Systematic care in designing representative questions is important to obtain valid information from question-answer comprehension tests.

Written Graded Reading Assessments

Alcock et al. (Alcock et al., 2000), working in Tanzania, developed a test of reading which aimed to be sensitive to early literacy skills and thus be suitable for administration amongst children who are just beginning to read. The tests were developed in a written format allowing for group administration and thus also for relatively inexpensive large-scale assessment. A further explicit aim of the test was to provide a valid measure of reading ability in languages with a shallow orthography (in this case Swahili) by focusing on recognition (of words) and comprehension (of sentences) rather than relying solely on fluency.

The assessment consists of three tests: of letter reading, word reading and sentence reading. The letter reading test required children to discriminate between letters and pseudo-letters, indicating their response with a tick or cross. Similarly, the word reading test required children to distinguish amongst real words, taken from first grade reading books, possible Swahili nonwords. The sentences reading task was based on the Silly Sentences task (Baddeley, Emslie, & Nimmo-Smith, 1992), This was a speeded comprehension task requiring children to indicate whether simple sentences were (Is your hand attached to your arm?) or false (Is your hand attached to your leg?). Children complete as many sentence as possible in five minutes.

In the original study the tests were found to be valid measures of reading ability when assessed against classroom achievement. Test retest reliability was reasonable for the letter reading task (0.77) and good for word reading (0.92) and sentence reading (0.89). Because of their potential for cost-effective group administration and their applicability to languages with shallow or deep orthographies, these tests were further piloted for inclusion in the World Bank multi-country impact assessment battery and results are reported below,.

Maze Test

A potential, alternate informal assessment of early reading ability and reading comprehension is the Maze test. The Maze is a multiple-choice variation of the Cloze test. Both the Maze and the Cloze test involve a passage where a few choice words are replaced by a blank. Children are required to read the passage and provide the missing word. While the Cloze format requires readers to generate the missing word, the Maze format requires readers to choose the target word from among three or four foils presented in a multiple-choice format. Hence, the Maze format addresses some of the drawbacks of the Cloze test. Chief among these are the high difficultly level of the Cloze format for beginning readers as it requires them to generate responses for each blank (Gillingham & Garner, 1992) and that it tests superficial skills and discounts reasonable alternate responses that are generated by students (Parker, Hasbrouck, & Tindal, 1992). However, it is important to note that on the technical front there is limited research evaluating the psychometric properties of the Maze test and on the substantive front there is some debate about the adequacy of the Maze test as an index of reading comprehension that adequately captures processes such as schema construction or inference generation. However as a broad measure of reading ability it is a reasonably useful measure, especially for large scale assessments as it is group administered and therefore time and cost effective.

In constructing a Maze test, the first sentence of the passage is left unaltered. From the rest of the passage, the target words can be deleted using one of two approaches: (a) a 'fixed ratio deletion' method where every fifth or seventh word is deleted or (b) a 'lexical deletion' method where selected words from specific classes - e. g. nouns or main verbs - are are deleted. The latter method, which is a more promising approach of the two methods, is based on the rationale that these classes of words contribute more to the semantic content and therefore provide a more valid index of passage comprehension than would the deletion of other classes of words. For example, correct selection of deleted words such as auxilliary verbs, or pronouns, is more likely to test syntactic competence than comprehension. Additionally, the lexical deletion method provides for cross-linguistic comparability by controlling for language specific syntactic factors.

The Maze test is administered as a timed measure (ranging from 1 to 3 minutes) so as to avoid scores that are negatively skewed as a result of most children getting all the items right. Timed versions of the Maze are also likely to have higher validity coefficients (Parker, Hasbrouck, & Tindal, 1992; Wiley & Deno, 2005). Scores on the Maze represent the number of correct word choices made in the prescribed time limit.

Other issues in the construction of a Maze test relate to the choice of foils or distractor words. The review of the limited number of studies that have used the Maze test (Parker et al, 1992) suggests that the selection of distractors should be from the same class of words as the target word, i.e. if the target word is a verb, then the distractors should also be verbs. If this is not the case, the test is more likely to be an assessment of grammatical knowledge rather than reading comprehension. Moreover, distractors should be of a similar level of familiarity as the target word, and the inclusion of three or four distractors along with the target words is likely to reduce the effect of guessing. Finally and most importantly, for a test of 'passage' comprehension it is important that the test be so constructed that it requires more than just comprehension at the 'one-sentence' level (Parker, Hasbrouck, & Tindal, 1992).

PILOT STUDY

A pilot study was conducted to develop and evaluate strategies for large scale multi-country assessment of reading ability. The study is described in full in Appendix A but findings are summarized here grouped according to principle objectives of the study.

1) To develop and evaluate a measure of reading fluency that is valid in shallow othographies and is a sensitive measure amongst poor readers.

Passage reading fluency was highly related to comprehension in both Swahili (a shallow orthography where fluency in the absence of comprehension is possible) and in English (Tables A4, A5 and A7 in Appendix A). We conclude that fluency measures are a valid method for cross-linguistic assessment of reading ability.

A composite of letter reading and passage reading fluency scores offers sensitivity across the whole ability spectrum (Figures A10 and A18) and is closely related to reading comprehension. Adding an intermediate level to this assessment battery – a test of non-word reading – adds little information to this measure and complicates test development considerably. We conclude that a combination of letter reading fluency and passage reading fluency is the most effective method for assessment of reading ability at the Grade 2 level in developing countries.

(2) To evaluate group-administered written tests of reading ability

This test consisted of three stages – letter recognition, word recognition and sentence reading. Piloting was conducted in Swahili only. Results suggest that the letter recognition test lacked validity. A revised version of the test battery including individually administered letter reading fluency test proved to be highly related to reading comprehension (Tables A4 and A5). We conclude that this combination of oral and written tests is a valid method for cross-linguistic assessment of reading ability. Compared to oral fluency measures (above) these measures have the advantage of being largely group administered. Their disadvantage is that test development presents greater challenges than for fluency measures. This battery of tests could be used where it is important to reduce cost of assessments on a very large scale.

(3) To evaluate the Maze measure as an efficient group-based means of assessing reading comprehension directly.

The distribution of scores on the Swahili version of the Maze test was acceptable (Figure A8) and the correlation with reading comprehension was high (Tables A4 and A5) suggesting the measure was sensitive across the ability spectrum and also a valid assessment of reading ability.

However, the English version of the Maze test proved very difficult for participating children. Most children failed to answer a question correctly (Figure A16)

and the correlation with reading comprehension was low (Table A7). That the test was more difficult in English than in Swahili highlights the problems of cross-linguistic comparability of this test for a number of reasons discussed above. It also emphasizes the difficulties in constructing this test to be sensitive across the ability spectrum. We conclude that this is not an appropriate method for cross-linguistic assessment of reading ability.

RECOMMENDATIONS

The recommended version of the reading assessment would consist of the following.

(1) An oral letter reading fluency tests – children read as many letters/graphic units as possible in 60 seconds.

(2) An oral passage reading fluency tests – children reading as many words of a connected text as possible in 60 seconds. Two different passages should be used to improve reliability.

(3) Comprehension should be assessed by 5 questions asked at the end of each passage (students who do not finish reading the passage in 60 seconds should be allowed to finish).

Outcome measure

The principle outcome measures reported will be the

(1) number of letter read per minute and

(2) the number of words read per minute (averaged from the two passages).

Comprehension questions should not be reported as an outcome measure – problems in generating questions of comparable difficulty between languages precludes the possibility of meaningful cross-linguistic comparisons.

For the purpose of analyses of the outcome measure (e.g. regression analyses or impact assessments) a single metric can be created. Results suggest that the two fluency measures (letter reading and passage reading) contribute to an overall score with equal weight (see Appendix A). The composite measure should therefore be the mean of standardized scores for the two fluency measures. It should be noted that it will be difficult to interpret the meaning of this composite measure so summary statistics should report the original fluency scores wherever possible.

Table 2 provides a list of the recommended assessments along with a brief description of the test development, required testing material, administration and training procedure. Table 3 provides a list of some of the commonly available commercial tests of oral reading fluency in English that have been developed and normed for use in the U.S.A.; Table 4 provides a summary of some of the studies of early reading ability.

NEXT STEPS

(1) Tests of oral letter and passage reading fluency should be developed in a number of countries. Two passages should be selected from the end of the Grade I text books in each country. Letter reading tests should consist of randomly selected letters from alphabets (or other collection of graphic units). In many languages, children learn the entire set of graphic units (the whole alphabet) from Grade 1. In other languages, such as Hindi, a subset of more common graphic units are learnt first. In this case we offer two suggestions for selecting this subset:

(i) the recommended subset presented in textbooks is used

(ii) letters/graphic units are taken from the chosen reading passages – if this latter strategy is used, the same procedure should be carried out for all languages

(2) Before testing begins, the text should be analyzed for length and complexity. First, the number of phonemes, syllables and words should be assessed in the passages in each language to test the hypothesis that cross-linguistic variability in number of phonemes is less than that for number of syllables and words. Second, readability

formulae should be applied to each passage. Primarily, a measure of lexical diversity should be applied to all passages and acceptable limits for this variable established. Passages falling outside the criteria should be replaced.

(3) Further pilot work could be conducted to establish the psychometric properties of the measures – principally the correlation between fluency and comprehension (a measure of validity), the correlation between fluency of reading the two different passages (a measure of alternative form reliability) and the sensitivity of the measure across the ability spectrum.

Table 2: List of recommended reading assessments

Type of Assessment	Testing Material and Development	Administration	Time Required	Scoring	Training
Letter Reading Fluency: to determine automaticity with the writing system, for e.g. letters of the English alphabet for children learning to read English[2]	Sheet with randomly printed letters and a stop watch.	Students are presented with a list of randomly arranged letters and asked to read aloud for one minute the names/sounds of the letters as fast as they can. Examiners note down the errors as students read aloud the name/sound of letters. If a student takes longer than 3 seconds to identify any letters, examiners encourage them to move to the next letter.	One minute of administration time and a few minutes to familiarize the student with the task requirements.	Letter reading fluency is scored as the number of correctly identified letters, either by their sounds or their names in a span of one minute.	Requires basic training of examiners to familiarize them with the standardized administration methodology.
Oral Reading Fluency: to determine accuracy and	Passages selected from grade level primers and a	Students are presented with a passage and asked to read aloud at a comfortable pace as	Two to three minutes of administration time and a few additional minutes to	Oral reading fluency is scored as the number of correctly read words in a span	Training of examiners involves familiarizing them with the standardized

[2] Note: For languages like English, both upper and lower case letters are included.

rate in decoding words in context.	stop watch Two passages are selected from the end of the Grade I reading textbook.	they will be asked questions about the passage. The examiner allows the child to read the entire passage and makes a note of the last word read at the end of a minute. The examiner also notes the errors. If the child hesitates for longer than 3 seconds on a word, the examiner provides the word and notes it down as an error	read the instructions to the student.	of one minute.	administration procedure as well as training to reliably note reading errors
Question-Answer Tests: to check students' comprehension of the passage	Four to five questions based on the reading passage. The questions tap literal information from the passage and are not inferential in nature.	Students are orally presented with the questions and their responses noted down.	Requires approximately one minute to administer each question. Hence, depending upon the number of questions it requires 4 to 5 minutes as well as a couple of additional minutes to read the instructions to the student.	Correct responses receive a score of 1 and scores range from 0 to 4 or 5 depending upon the number of test questions.	Training of examiners involves familiarizing them with the standardized administration procedure as well as with the potential correct responses to the test questions.

Table 3: A summary list of some of the commonly used commercial assessments of oral reading fluency (ORF)

Assessment	Publisher	Description	Technical Adequacy
Dynamic Indicators of Basic Early Literacy Skills (DIBELS)	University of Oregon Center on Teaching and Learning	It is a standardized, normed, and individually administered measure of ORF in connected text designed for grades first to third. Children are asked to read aloud a text for one minute and oral reading fluency rates are reported as number of correct words read per minute. An accompanying 'retell fluency' measure is administered to assess comprehension.	Test-retest reliabilities range from .92 to .97 Alternate form reliability of different passages drawn from the same level range from .89 to .94 Criterion-related validity coefficients based on eight studies range from .52 to .91
Test of Reading Fluency	Children's Educational Services, 1987	To include	To include
Reading Fluency Indicator	AGS Publishing	Is a standardized, normed, and individually administered measure of oral reading rates, and accuracy for grades one to twelve, and ages five to eighteen. The test provides four passages of similar difficulty for each level. Administration time varies from less than one minute to several minutes depending on the amount of time a child takes to read	Not provided

		the passage. Oral reading fluency rates are expressed as number of correct words read per minute. The reading passages are also accompanied by four comprehension questions.	
Reading Fluency Progress Monitor	Read Naturally	Is a standardized, normed, and individually administered test of oral reading fluency. It comprises grade level passages for grades one to eight. The availability of 30 passages for each grade level allow for monitoring individual student's performance. Students are made to read aloud for one minute and their oral reading fluency scores are reported as correct words read per minute.	
Reading Fluency Benchmark Assessor	Read Naturally	To include	To include
Gray Oral Reading Test, Fourth Edition (GORT-4)	PRO-ED	For ages 6-18. The test comprises two parallel forms, each of which have fourteen developmentally sequenced reading passages with five accompanying comprehension questions.	Internal consistency reliabilities on average are .90 or above.

Table 4: A review of large scale, single and multiple country reading assessment studies. The focus of this review is chiefly on test development and administration procedures

	Participating Countries	**Aim**	**Age Group**	**Methodology**	**Reported Findings**
The use of national reading tests for international comparisons: ways of overcoming cultural bias, Bonnet et al (YEAR??)	England, Finland, France, Italy	A feasibility study to test the use of an alternate testing methodology that involves the use of indigenous testing material as opposed to translations of one common test in order to avoid bias resulting from linguistic and cultural factors.	15 year old	Each country developed its own assessments based on authentic literature. Common items from the WISC-III verbal subtest were administered to anchor the assessments across the countries Bilingual students from each country to enhance comparability	The methodology is promising warranting further development, especially for: (a) selection of type of text, (b) selection of type of questions (multiple-choice or open-ended), and (c) coding scheme The use of an anchor test is essential
Progress in International Reading	Thirty five countries	PIRLS is an international	The target population in	The selection and translation of assessment passages and the	The findings indicate substantial

Literacy Study (PIRLS): International Association for the Evaluation of Educational Achievement		evaluation study of reading providing extensive and comparative data on reading achievement based on a rigorous sampling and testing methodology.	most countries is 4th grade students as the idea was to assess children who had gained mastery of basic reading skills and were more focused on "reading to learn".	development of test items, and scoring guides was based on a rigorous and intensive process spanning over two years. The assessment of reading ability is based on constructed-answer responses and multiple-choice items and requires extensive testing time (80 minutes for each child to complete two passages) due to the large set of test items. For the evaluation of constructed-response answers detailed guides and rubrics were developed as well as intensive training provided for the team of raters.	differences in reading ability for fourth grade students across and within countries and examines key home and school contextual factors associated with these differences.
Fourth-Grade Students Reading Aloud: NAEP 2002 Special Study of Oral Reading (National Center for Education Statistics (NCES))	U.S.A.	A sub-study of the NAEP to evaluate the oral reading ability of fourth grade students via direct observation of reading to supplement and get a more complete understanding of paper-pencil	A nationally representative sub-sample of 1, 779 fourth grade students drawn from the main NAEP reading assessment.	Students, screened on a second-grade level passage, were asked to silently read a grade level passage. Students were then asked three constructed-response comprehension questions based on the passage. This was followed by reading aloud of a section of the passage. This reading was timed and recorded. Recordings of student readings of grade level passages were analyzed by trained raters for fluency and accuracy.	All the three measures of oral reading ability – accuracy, rate, and fluency were positively correlated to each other and all three were positively associated with reading comprehension. Error analysis indicated that

		tests.		*Fluency* defined as "phrasing, adherence to the author's syntax, and expressiveness" was rated on a scale from 1 to 4. Accuracy was coded based on number of errors made by students. Reading *rate* was reported in two ways: (a) number of words read in the first minute of reading, and (b) number of words per minute for the entire passage.	reading errors that changed the meaning of the text were more directly related to reading comprehension than were errors that did not result in change of meaning.
Annual Status of Education Report (2005): Pratham, India	India	A large scale national survey to assess reading and numeracy skills of primary grade students attending government schools in rural India	A national sample of primary grade students attending government schools in rural India sampled from 485 of India's 603 districts	All children are presented with simple text comprising four sentences to read. Based on their performance on this text they are presented with either an advanced passage or a list of words to read. Basic guidelines are provided to all examiners to rate students reading levels. For instance *fluency* when reading a passage is evaluated on the basis of the following guidelines: "If she can read fluently with ease and reads like she is reading a long text, then she is marked as a "story"	Findings reported by grade and state level indicate dismal reading performance levels of children attending government schools in rural India.

	child. This child can read LEVEL 2 text.”

Review of Classroom Observation Assessments

OVERVIEW OF CLASSROOM OBSERVATIONS

Introduction

Assessment of teaching and learning processes in the classroom is an essential part of any attempt to monitor or evaluate the quality of education {e.g. \UNESCO, 2005 #2281} The key method for evaluating the nature of instruction and teacher/student behavior involves direct observation of teachers and students in the classroom. The classroom observation is a relatively affordable means for obtaining objective and quantifiable records of teacher and student behaviors in classroom (Medley, 1982). It has been used for various purposes such as program evaluation, examination of fidelity of program implementations, and examination of the instructional practices that encourage higher level thought processes (Hillberg, Waxman, & Tharp, 2004). In particular, starting in the mid-80s formal classroom observations have been used in many studies of school and teacher evaluation research, particularly in the US context (Ellett & Teddlie, 2003).

Approaches to Classroom Observation

There are different types of classroom observations:

(1) systematic observations;
(2) narrative description (including ecological observations, ethnographic observations);
(3) judgment-based ratings of teacher characteristics.

These different methods not only differ in format but also serve different purposes. For instance, narrative description can be useful as a discovery tool for classroom behaviors and provides “thick” description on classroom interactions. However, narrative description tends to be more vulnerable to subjectivity and processing and analyzing the

narrative description data becomes a challenge in a large scale cross-country comparison studies. In contrast, systematic observations offer some advantages such as providing reliable (given rigorous training) and quantifiable data, and they can be used repeatedly over time. Also the results can be easily shared and communicated among those who are familiar with the system (Bramlett & Barnett, 1993). Thus, we focus on the systematic observation in this report.

Systematic observation of classrooms refers to "observations of classroom behavior made by a trained observer who records the behaviors according to an observation system." (Medley, 1982). It collects quantitative observational data for direct observations. Systematic observation system consists of a list of prespecified items or categories of behavior to be observed, thus behaviors not listed are ignored (Medley, Coker, & Soar, 1984). Systematic observation collects data on frequency and length of specific behaviors in the classroom. Events or behaviors to be observed and methods to record them are prespecified and observers usually tally or mark the occurrence or frequency of preidentified behaviors during an observation unit (e.g., every 30 seconds). Generally the categories for predetermined behaviors are extensive, exclusive, and well defined (Boehm & Weingberg, 1997). Systematic observation instruments involve the following: (1) an explicit aim of the observation, (2) operational definitions of all observed behaviors, (3) training procedures, (4) a specific observation focus, (5) setting, (6) a unit of time, (7) an observation schedule, (8) a method to record the data, and (9) methods to process and analyze data (Stallings & Mohlman, 1988).

Items in observational instruments vary in the degree to which inference is required. For example, items that have a rating scale require high inference whereas checking the frequency of a certain behavior requires low inference. There are three basic types of systematic observation systems: category systems, sign systems, and multiple coding systems (Medley, Coker, & Soar, 1984). Category systems have a set of behavior items or categories, which are defined broadly so that observed behaviors should be classifiable as an instance of one or another category in each domain. The sign system, on the other hand, has a list of narrowly defined behaviors (signs), which are relevant to a dimension

of behavior to be measured. A multiple coding system codes a single behavior or event into multiple categories. For example, the Stallings Observation System includes a time sampling process in which an individual interactive statement (during a given time, e.g., five minutes) is coded in four different categories: (1) Who?, (2) To whom?, (3) What?, and (4) How?

Diversity of classroom context, cultural expectations across countries

Countries vary greatly in macro-level factors (e.g., state policies, international organizations) and micro-level factors (e.g., school structures, community life, and family socioeconomic status) that make cross-country comparison of classroom instruction challenging. In addition, education system and classroom practices are embedded in larger cultural systems such that classroom practices vary according to cultural beliefs, expectations, and practices. Therefore, simple application of an observation system that was created for developed countries may fail to capture intricate and important aspects of classroom behaviors in less developed countries. In order to capture the nature of classroom instruction appropriately, the systematic classroom observation instrument needs to have an extensive array of prespecified teacher and student behaviors to capture variation in teacher behaviors, student response, and teacher effectiveness in *diverse* cultural contexts. However, there is no one measure that will meet the needs of different context in different countries. Thus, the focus in this report was to find and recommend instruments that include key features that were reported to be important for children's learning in the classroom and/or instruments that are flexible to be adapted to different research purposes.

KEY COMPONENTS IN CLASSROOM OBSERVATION PROTOCOLS

Introduction

Despite the popularity of the classroom observations for its role in providing information for teacher and student behaviors, there are many caveats and complexities that need to be addressed in the selection and use of a classroom observation system. First of all, to have confidence in the conclusions drawn from the observational data, we need to be confident that the observational data reflect what really happened. This requires sufficient evidence for reliability and validity. Furthermore, classroom observation protocols should be easy to understand and use. Observation measures should also contain key instructional features and classroom behaviors that have proven to be essential for effective teaching and learning based on thorough literature review. Below is a brief discussion of essential requirements of systematic classroom observation: reliability, validity, ease of use, and substantive components of observational protocols (i.e., the use of instruction time framework in this report). For a reference, a short list of checkpoints for determining the appropriate observational system summarized by Boehm and Weinberg (1997) is found in Appendix A.

Reliability

It is critical to achieve agreement and consistency on the observational bases. Observational data are inherently vulnerable to subjectivity, which is usually influenced by observers' gender, race, age, bias, and expectations. For example, observers' beliefs and prior experiences or knowledge can lead to misinterpretation of what they observe instead of describing what really happened objectively (Good & Brophy, 1994). Thus, it is critical for classroom observation protocols to be reliable across observers (inter-individual agreement) and across time (intra-individual agreement) (Boehm & Weinberg, 1997). In other words, when two people observe in the same situation with a focus on the same behaviors, they should agree about what they observe. Also when the observer is in the same situation in different times with a focus on the same behaviors, s/he should be consistent in his/her description of the situation.

High rate of agreement and consistency among observers can be achieved through systematic and rigorous training. In addition, the amount of subjectivity is reduced when behaviors to be observed are precisely and unambiguously defined. Thus, it is imperative for an observation system to have clear and precise operational definitions for the prespecified target behaviors to observe and record.

Validity

In order for conclusions and inferences from observational data to be valid, the observation system should sample or represent the behavior of the concern. The content of an observational system should accurately reflect the conceptual question investigators want to address. The observation categories or codes should cover the target behaviors exhaustively and should be refined sufficiently. Furthermore, it is important for an observation protocol to have concurrent and predictive validity evidence. That is, observed behaviors are found to be related to concurrent and future related criteria/behavior. Finally, reactive effects of having an observer in the classroom may raise a concern about the generalizability of inferences drawn from observational data. In other words, one may question whether observational data reflect typical classroom process and interactions since observations are anxiety producing for teachers (Good & Brophy, 1994) and the presence of an unfamiliar adult may affect student behaviors. This is of particular concern when observations are high stakes as part of teacher performance reviews. However, it was demonstrated that when teachers are aware of and assured of the low stakes nature of classroom observations, teacher behaviors tend to be consistent from one observation occasion to another (Tollefson, Lee, & Webber, 2001).

Ease of Use

User-friendliness is another important aspect of an observation system: The instrument needs to capture what needs to be observed (both teacher and student behaviors) in a manageable way (e.g., a fairly short list). A complex observational system not only takes more extensive training of observers but also is more vulnerable to inconsistencies among

observers in what and how to code what is observed. Many complex categories and codes demand great amount of memory which makes the implementation of the observation system more challenging and more prone to inconsistency among observers. Therefore, clear, nonoverlapping, and manageable number of categories, codes, and questions are essential for ease of use as well as for reliability, particularly in a large scale, international studies.

Key Substantive Features in Systematic Observational System: Use of Instructional Time

Research suggests that school effectiveness plays a large role in student learning and achievement. School effects have larger influence on students' achievement than the family background in the less developed countries (Fuller et al., 1999; Heyneman & Loxley, 1983). Specifically, 81% to 90% of the total variance in student achievement in science was due to teachers and schools in less developed countries (i.e., Thailand, Brazil, Colombia, and India) while only 22 to 27% of variance was ascribed to teachers and schools in developed countries (e.g., Sweden and Australia).

There are multiple indicators of school effectiveness (e.g., school management/leadership, and teacher quality), but in this report, we focus on teacher quality, teachers' time management in particular, based on the influential Carroll (1963) Model of School Learning as well as teacher effectiveness research literature (e.g., Brophy & Good, 1986). The teacher effectiveness literature indicates that loss of instructional time is related to student achievement and a major impediment on improving instructional quality. For instance, the loss of instructional time due to teacher absenteeism contributes to loss of student learning opportunity, and thus low student achievement. Furthermore, teachers' effective use of fixed instructional time promotes and maximizes student learning. Below we provide a brief literature review on the instructional time framework in the examination of school and program effectiveness.

Allocated instructional time and loss of instructional time

Instructional time is a necessary and pivotal condition for learning to occur (Avot, 2004; Carroll, 1963; Stallings, 1980). Children need sufficient exposure to learning materials in order to successfully acquire target materials. An easy way to measure is obtaining information about allocated instructional time. However, substantial and pervasive discrepancies have been observed between intended/allocated time and school realities in instructional time, particularly in less-developed countries. For example, in Ghana, many rural school teachers were reported not to follow the prescribed weekly timetable (EARC, 2003, cited in Avot, 2004).

The loss of allocated instructional time is attributed to many different factors at various levels. For example, the lack of government's ability to reinforce the full implementation of official time guidelines is a source in many less-developed countries. At the school level, the loss of instructional time is attributed to poor physical condition and poor infrastructure (e.g., high noise level & lack of heating) and lack of teachers. The loss of instructional time also occurs due to the scarcity of school resources. For example, in Gambia and Burkina Faso a large chunk of instructional time was spent on writing lessons and problems on the board due to students' lack of access to textbooks (Avot, 2004). At the classroom level, teachers lose instructional time for various reasons such as poor teacher training, high levels of teacher attrition, overloaded curricula, and the school teachers' low social position.

Teacher absenteeism is one of the major sources of loss of instructional time, in less-developed countries in particular. Studies revealed that teacher absenteeism varies from 20 to 30 percent in less-developed countries on a given day, on average. In Pakistan, about 18 percent of public and private school teachers were absent (Ali & Reed, 1994). In India about 33% (PROBE, 1999) and 37% (World Bank, 2001) of teachers were surveyed to be absent from classrooms. In Indonesia, 21 percent of teachers were absent from school and 27 percent of those at school were out of classrooms (Rogers, 2003). Moreover, the teacher absenteeism is even higher in rural schools (Baker, 1988; Pitkoff, 1993) and double-shift schools (Linden, 2001).

Student time-on-task and distribution of activities across time

Research suggests that how available instructional time is used, not just the amount of time available, is the key to successful instruction and student learning. Creating students' opportunities to learn and increasing pupil's engaged time in learning is the ultimate underpinning of students' academic success. The critical element is maximizing opportunity-to-learn/time-spent-in-learning by using fixed amount of instructional time effectively. Previous studies demonstrated that teachers' use of instructional time is intricately associated with instructional quality. Teachers' use of time (or the distribution of time across activities) has an impact on student achievement (Good & Brophy, 2000). A study in Pakistan revealed that teachers' effective use of time was a better predictor of student achievement than teaching time (Reimers, 1993). Tan and his colleagues (1997) and Verwimp (1999) also reported a similar finding in the Philippines and Ethiopia. In contrast, studies in Brazil revealed that substantial amount of class time was spent on students' mundane copying work while teachers graded papers and this kind of activity was negatively associated with their reading scores (Fuller et al., 1999).

The teacher's job is to organize and present materials to learners for efficient and rapid acquisition of the presented materials (Carroll, 1963). Teachers' use of instructional time has been used in teacher and school effectiveness studies in the US and international studies (e.g., Anderson, Ryan, Shapiro, 1989; Fuller et al., 1999; Goodlad, 1980; Stallings & Kaskowitz, 1974; Stallings & Freiberg, 1991; Teddlie, Kirby, & Stringfield, 1989). Teachers' effective use of active and passive instruction affects student learning (Stallings, 1980) and reduces the proportion of student off-task behavior (Brophy & Good, 1986; Reynolds, 1992; Waxman & Padron, 1995; Waxman & Walberg, 1991). Active instruction includes teacher instruction, questioning, providing supportive corrective feedback while passive instruction include judiciously planned seatwork such as review of the materials. It was shown that more time spent on discussion and review of the materials, more time on reading aloud, and on supportive, corrective feedback from the teacher were positively associated with student achievement (Stallings, 1980). In

contrast, more time spent on classroom management/organization and teachers' social interactions were negatively associated with student achievement (Stallings, 1980). Stallings report that the following proportion of the time spent appears to prove improved student achievement: at least 50% of teacher activities be active (or interactive) instruction, 85% the combined active and passive (or noninteractive) instruction, and less than 15% for organizing and management related activities. In addition, Smith (2000) found that the effect of loss of instructional time is profound in urban schools that serve disadvantaged and low-performing students.

Outside of the US context, Fuller and his colleagues (1999) examined the relationship of family characteristics, and teacher and classroom characteristics to children's literacy achievement in Brazil and found that teacher's use of time was closely related to children's literacy skills after controlling for their family background characteristics. Variation in teacher practices (mostly use of instructional time) added about 9% of unique variance in children's early reading skills after controlling for their family characteristics. Furthermore, a study examined and compared proportion of instruction time that is spent on different activities (using the Stallings Observation System) in Tunisia, Morocco, Brazil, and Ghana[3]. The study revealed that students in Ghana were offered 10 to 20 percent less instructional time and teachers in Ghana spent higher percentage of time on organization/management than the other countries.

[3] World Bank four country report on Classroom Behavior

DESCRIPTION OF SYSTEMATIC CLASSROOM OBSERVATION MEASURES

Introduction

Below we provide brief descriptions of systematic classroom observation measures that we surveyed. Table below presents a summary of the measures described in this section.

Stallings Classroom Observation System

The Stallings Observation System, technically called the "Stanford Research Institute Classroom Observation System", was developed in order to evaluate the implementation of several educational programs participating in the Follow Through Planned Variation Project (Stallings, 1978; Stallings & Kaskowitz, 1974). It was designed to be "broad enough or flexible enough to accommodate the wide range of projects". The Stallings Observation System is a low-inference system that measures teachers' and students' classroom behaviors, instructional strategies, and classroom environment. It has been widely used in the US context with various student population (kindergarten through 12th graders) and in variety of context of schools (e.g., urban, suburban, and rural) (Knight, 2001) and a few studies outside the US context. Studies have validitated the Stallings Observation System against measures of school effectiveness: consistent mean differences in teaching behaviors between schools were associated with effective, and less effective school characteristics (Ellett & Teddlie, 2003; Stallings, 1975; Stallings, Needels, and Stayrook, 1979). Finally, the Stallings system is available in various formats: paper-pencil, scantron, and a computerized system. The computerized Stallings Observation System makes data analysis easy for a large project, and can create database of teacher profiles. However, the Snapshot can be easily implemented by scantron or paper and pencil.

The Stallings observation system consists of three main instruments: (1) the Classroom Snapshot, (2) the Five-Minute Interaction, and (3) the Physical Environment Information. Below is a description of each of the instruments.

Stallings Classroom Snapshot **(**Stallings, 1980)

Stallings Classroom Snapshot has been used in many studies in school and teacher evaluation studies. The Snapshot provides a profile of teachers' and students' behaviors at different times. In a grid format it records every person (adults and students), the activities s/he is engaged in (e.g., reading, social interaction, student uninvolved, discipline, arts/crafts), materials used for each activity (e.g., textbooks), and whom s/he is engaged with (i.e., grouping configurations of students such as individual student, small group, large group, and whole class). It also provides information on the percent of time in which teachers and students are engaged in various activities, including off-task activities.

When using a paper version of the Stallings Snapshot, the observer records prespecified teacher/adult behaviors in the classroom and materials used during instruction. Appendix B displays an example of the adapted Snapshot for the World Bank Project (2003). On the page, possible instructional activities are listed at the left in each row of the table. These instructional activities include receiving assignments, reading, instruction/demonstration, written assignment/seatwork, practice/drill, discussion, kinesthetic, projects, social interaction, student uninvolved, being disciplined, and classroom management. Materials used for each of these activities are indicated in columns, including books, notebook/slate, chalkboard, computers/calculators, manipulation, visual aids, cooperative learning, and none. In addition, there are a set of letters to the left of each materials column, T, O, I, which stand for the category of participants in the classroom: teacher (T), other adult (O), and independent student (I). Another set of letters is under each materials used column, I, S, L, and E, to record grouping configurations for each activity: I (one student), S (more than one student, up to 30% of the class), L (more than 31% but less than the total group), and E (everyone).

The Snapshot is usually completed five times during a class period and each classroom is observed on at least two consecutive days yielding a total of ten Snapshot observations (Knight, 2001). The data are usually summarized into three composite variables for

teacher activity: monitoring seatwork, providing interactive instruction, and organizing/managing.

Some of the advantages of Stallings Classroom Snapshot are;

- The Stallings classroom snapshot provides information that is objective and the categories are easy to understand and use. Thus it is efficient in training, requiring just two half-day training sessions to yield reliability estimates above .90 (Knight, 2001).
- There are many studies that provide criterion validity evidence. The Snapshot has demonstrated consistent associations between teaching behaviors across schools and effective, typical, and less effective school classification (Ellett & Teddlie, 2003).
- It is easy to modify the categories of the types of classroom activities and materials used to better suit the need of a particular study or a context. For example, a few items of the Snapshot appended in this report are modified from the Snapshot used in the US context in order to reflect classroom context in less developed countries. Specifically, copying of materials off of the blackboard is a frequent activity in the classroom in less developed countries, which is not one of the categories in the Snapshot used in the US context.

Limitations of Stallings Classroom Snapshot include

- Its lack of capacity to capture the type and quality of the interaction and content of the lesson. Information about the quality of interaction can be obtained from Stallings Five-Minute Interaction Instrument (see below). However, the Stallings Five-Minute Interaction Instrument requires extensive training of observers.

Stallings Five-Minute Interaction Instrument

This instrument is developed to capture all the interactions among teacher and students in the classroom to describe and quantify the teaching style and the interaction processes in the classroom. Every statement/event made during the classroom observation is recorded in a series of four-celled the prespecified categories in sets of five-minute time blocks. Appendix C presents an example of Five-Minute Interaction frame (on an actual observation sheet, there are multiple frames). All interactions become coded in four different categories: (1) '*Who* does the action?,' (2) 'To *Whom* is it done?,' (3) '*What* is done?,' and (4) '*How is it done?* These categories identify the speaker, the person being spoken to, the message being delivered, and the emotional affect for each event/statement during observation. Under each of these four categories, there are a number of codes to be checked off. The Who? and To Whom? categories have 10 codes (e.g., teacher, aide, child, small group, animal, and machine) while the What? category has 14 codes (e.g., command, request, response, instruction, etc). The How? category records the tone of interaction in 12 codes (e.g., happy, unhappy, negative, etc.). The four categories are presented in a series of frames, each statement/event being coded in a frame. For example, a teacher's statement to a student, "What is two minus one?", would be coded as (1) *Who?* Teacher (2) *To whom?* Child (3) *What?* Request (4) *How?* Question.

The Five-Minute Interaction is completed five times distributed evenly across the class period. As in the Snapshot, data are collected for two consecutive days of observation for a total of 10 Five-Minute Interaction observations.

This instrument takes more time for training and is more prone to inconsistencies among observers because determining the nature of interactions requires more subjective judgment than the Snapshot. It is reported that the Five-Minute Interaction observation requires five days to learn to use the coding system to yield 85% or greater interrater agreement (Knight, 2001).

Physical Environment Checklist

The physical environment checklist records seating patterns and arrangements of instructional materials, play equipment, audio-visual equipment, general equipment and materials. It should be noted that Stallings (1980) deemphasized this part of the Stallings Observation System.

Extra Teacher Project (ETP) Observation Instrument

The ETP observation was designed to capture changes in classroom behavior following the provision of an extra teacher to reduce class sizes in the lower primary grades in Kenya. The content of this instrument is tailored to this specific goal, but the methodology includes innovations with general applicability to projects in developing countries.

The ETP observation instrument builds on the methods of the Stalling Observation System by addressing two problems of large-scale evaluations in developing countries. First, teachers behave differently when they are being observed. This reactive effect of observation affects *some* aspects of teaching more than others. For example, even when teachers make a conscious effort to improve teaching methodology when observed, they may lack the knowledge or regular practice to adopt different teaching methodology. In these circumstances, observed classroom behavior is more likely to reflect general teaching practice. However, other behaviors are more influenced by observation. Teachers are unlikely to spend time outside the classroom or on non-instructional activities (e.g. doing administrative work) when they are being observed. Such behaviors are more common in developing countries and are a key predictor of poor academic achievement (discussed above). The second problem is that observation measures require a number of visits to the school making large-scale assessment costly.

Two modifications were made to methodology in order to address these two problems

- Observations were conducted on surprise visits to schools to reduce possibility of teachers' planning for the assessment.
- Retrospective information was collected through teacher and pupil interviews and through examination of student exercise books. This provided some objective information, free from the effects of an observer, and also helped maximize the amount of information collected from a single visit.

The ETP observation system consists of Teacher Questionnaire, Student Questionnaire, and Observation Schedule. *The Observation Schedule* includes the coding of prespecified behaviors in fixed periods of time (e.g. 30 seconds). In addition, an overall assessment of the amount of time and type of activities in the lesson as a whole is provided. *The Teacher Questionnaire* includes questions about the lesson that took place immediately before the observers arrived at school. Questions concern teaching methods used and other classroom behaviors. *The Pupil Questionnaire* is group administered and orally presented to children while the interviewer completes the questionnaire. The observer checks off prespecified behaviors. The first section of this questionnaire concerns the lesson that took place immediately before the observers arrived (the same lesson that is the subject of the Teacher Questionnaire). Students are asked similar questions to the teacher about teaching methods used. They are also asked to report how much of the time the teacher was present. In the second section of the question 10 randomly selected pupils describe the nature of any interaction they had with the teacher in the previous lesson. In the final section the interviewer examines five randomly selected students' exercise books, looking at work completed in the previous four weeks. The examiner records the number of exercises students completed and the number that received written comments or grades from the teacher.

Data are not yet available on this instrument but experience of implementing this assessment suggests that pupils, even in Grade 2, can provide detailed information about previous lessons. Data analysis (taking place early in 2007) will assess the accuracy of this information. Analysis will also provide other information on reliability and validity of the overall instrument. Experience also suggests that surprise visits are an effective

way to capture valid information about classroom behavior. However, this entails careful attention to maintain the relationship between the schools and the organization conducting the observations, which is crucial. In the Extra Teacher Project, co-operation from schools was improved by explaining that the study was independent from the Kenyan government and that information be anonymized. On the other hand, as combination of the unscheduled arrivals and the perceived lack of governmental authority behind the study meant that schools occasionally requested the observer teams to return on a scheduled date, contradicting the aim of unannounced visits. The ETP observation system can be found in documents accompanying this report.

Special Strategies Observation System – Revised (SSOS-R, Schaffer, Nesselrodt, & Stringfield, 2004).

The SSOS-R is a comprehensive observation system based on previous observation systems and the learning and teaching model (e.g., the Stallings Observation System, the Classroom Activity Record, and Slavin's QAIT model). SSOS-R consists of three instruments: QAIT Assessment of Classroom, Classroom Observation Schedule, and Classroom Environment and Resources Checklist. It includes both quantitative and qualitative information from classroom observations: Classroom Observation Schedule records frequencies of different teaching behaviors while QAIT provides "thick" descriptions of the contexts within classrooms (Schaffer, Nesselrodt, & Stringfield, 2004). Several studies have shown reliability and validity evidence for this system (Meehan, Cowley, Finch, Chadwick, Ermolov, & Riffle, 2004; Nesselrodt & Schaffer, 1993; Schaffer, Nesselrodt, & Stringfield, 1991; Stringfield, Winfield, Millsap, Puma, Gamse, & Randall, 1994). However, because the whole system is extensive and comprehensive, it requires extensive training of observers in order to achieve acceptable reliability. The SSOS-R can be optically scanned or administered in paper and pencil.

QAIT

The QAIT stands for Quality of instruction, Appropriateness of instruction, Incentives for learning, and Time involved in learning. The four categories were derived from effective teaching and learning practices, based on Slavin's synthesis of literature on effective learning and teaching (1987, 1989). It is a high inference, simple coding, rating instrument and is completed at the end of classroom observation. It is a rating instrument with a 5-point Likert like scale (1 for 'Unlike this class' 5 for 'Like this class'). It has 40 items classified under four categories on a 8 ½ by 11 sheet. Coding of time-on-task in the QAIT instrument includes whether necessary time is allocated for instruction and student are engaged through teachers' effective management skills. The QAIT assessment usually occurs three times during an hour-long observation. The QAIT instrument, the reliability coefficient for all 40 items was .96, the coefficients for each of the scales ranging from .69 to .95 (Hughes, Cowley, Copley, Finch, & Meehan, 2005). Although this instrument, as in any other instruments that employ rating, is subject to higher rate of subjectivity compared to low-inference observation instruments, this instrument provides information on the *quality* of instruction. For acceptable reliability, it appears that it will require extensive training of observers. However, no published information is available about the length of training of observers.

The Classroom Observation Schedule

The Classroom Observation Schedule has low-inference items with multiple coding procedures, based on the Stallings Observation System (Stallings, 1980) and the Classroom Activity Record (Evertson & Burry, 1989). The Classroom Observation Schedule allows observer to code the entire classroom and also focus students (up to 3 students). This observation system takes 58 minutes to complete including 2 minutes for cover page for demographic information and 56 minutes which are divided into seven 8-minute time periods; each 8-minute block is captured on a separate page. The 8 minute of observation consists of recording student engagement and grouping strategies for the first minute while the rest of seven minutes are spent on the target student(s). The Classroom Observation Schedule includes 'classroom activity codes', 'student

engagement rate', 'group configurations,' and 'time spent.' For classroom activity codes there are 27 specific teaching or student activities under three different categories: (1) Teacher led, (2) Student/Group led (e.g., sustained writing, reading; independent inquiry or research), and (3) Management/Organization. These 27 activity codes are repeated six times so that a total of six different activities can be coded during the eight-minute time period. Student engagement rate is measured by recording (a) number of students on task, (b) number of students off task, (c) number of students out of room, and (d) number of students waiting. Group configuration information during activities is captured in four categories: interactive instruction, work alone, management/directions, and social/uninvolved. Finally, the length of time spent for each activity is recorded.

The reliability coefficients for the Classroom Observation Schedule are as follows: .51 for the activity code section coefficient, .76 for the student engagement, and .76 for the grouping section and .38 for the number of students section (Hughes et al., 2005). It should be noted that the reliability coefficients for a couple of sections are low.

The Classroom Environment and Resources Checklist

The Classroom Environment and Resources Checklist is completed at the end of the observation period. It has 12 environmental items and 22 resources items that are coded either as present or not present. The reliability estimates for the Checklist for all 50 items was .87; the coefficients for each section ranged from .44 to .81.

Virgilio Teacher Behavior Inventory (VTBI, Teddlie, Virgiio, & Oescher, 1990)

Virgilio Teacher Behavior Inventory was developed to assess teacher effectiveness, based on a review of the research literature on teacher effectiveness (Virgilio, 1987). It provides information on three major areas: teachers' classroom management, quality of

instruction, and social psychological climate. These three skills areas for teacher effectiveness are captured by 35-items on a six point Likert format scale. The scale points of 1 to 5 corresponds to “poor”, “below average”, “average”, “good/above average”, and “excellent” while a point of 6 indicates not applicable or not observed. It has a high reliability estimates (ranging from .85 to .96 for three subcategories) and it is significantly correlated with a time-on-task measure, the Stallings Snapshot ($r = .64$, $p < .0001$), providing evidence for concurrent validity. Unfortunately, no published information on the extent of training of observers is available.

Classroom Observations Keyed for Effectiveness Research (COKER, Coker & Coker, 1979)

COKER is a low-inference sign system for observing teacher and student classroom behaviors, including teacher and student interactions, affective behaviors, and grouping configurations. The total observation time per visit is 10 minutes (two 5 minute observations using forms, Section A and Section B) although the actual time spent in the classroom will be longer. Within each observation period, the observer records two sections of the COKER, Section A & Section B. Section A records whether prespecified teacher and student behaviors are present or not present. The teacher behaviors have three categories: initiating behaviors, presenting, and responding. Each of these categories has multiple codes: Initiating behaviors have 13 codes, presenting behaviors have 7 codes, and responding behaviors have 14 codes. The student behaviors include 7 on-task and 3 off-task behaviors. Section B of the COKER system records teacher and student affective behaviors (both verbal and nonverbal) and teaching strategies observed during the five-minute observation period. The observer completes Section B from memory after the five minute observation period.

The reliability and validity information, and the length of training for COKER is not readily available.

Spaulding Teacher Activity Recording Schedule (STARS, Spaulding, 1982) & Coping Analysis Schedule for Educational Settings (CASES, Simon & Boyer, 1967).

The STARS and CASES are used in conjunction to provide a "snapshot" of the classroom interactions (Simon & Boyer, 1967). The STARS is designed to record the cognitive instructional methods, affective relationships, and behavior control strategies of teachers (Medley, Coker, & Soar, 1984) while the accompanying the Coping Analysis Schedule for Educational Settings (CASES, Simon & Boyer, 1967) codes student classroom behaviors. When used with CASES, 475 possible categories of data can be obtained (19 CASES categories X 25 STARS categories).

STARS

The STARS provides information on 25 teacher behavior patterns found commonly in the US preschool, elementary, and high school classrooms. These 25 behavior patterns are subsumed under 8 categories such as general, affective, motor or social structuring, concept formation and development, motivation and focusing, concept checking, valuing, and listening and observing.

Training of observers takes two to three weeks, yielding reliability estimates above .80 (Simon & Boyer, 1967). Data are usually taken by means of time sampling techniques on a 10-second sampling schedule. This yields information on frequency of various teacher and student behaviors.

Reliability of observers ranged from .70s to .90s depending on the complexity of the observed classroom settings. Also the generalizability analysis yielded a correlation coefficient of .80 using four observations, and .90 using eight observations (Spaulding, 1982). Five of the 25 behavior categories have been evaluated for validity (Weinrott, Jones, & Boler, 1981) showing convergent and discriminant validity.

CASES

The CASES records student behavior patterns in 19 prespecified categories, which were developed based on ego theory and personality development (Medley, Coker, & Soar, 1984). It provides useful information on process of students' socialization and feedback to teachers on the effectiveness of classroom management and instructional strategies (Simon & Boyer, 1967). Out of 19 categories of student behaviors, 13 categories describe coping behaviors identified by descriptive statements (e.g., aggressive, hurtful, destructive behavior, resisting, delaying, and defensive checking) and additional six subcategories code child behavior according to adult and cultural expectations (e.g., controlling others in pro-social manner or in a self-serving manner). Each observation takes about 15 to 20 minutes. CASES data can be taken continuously or time-sampled (e.g., every 10 seconds) (Simon & Boyer, 1967). The training of observers for the CASES takes about 25 to 30 hours to yield an acceptable reliability above .80 (Medley, Coker, & Soar, 1984; Simon & Boyer, 1967).

Reliability of CASES yielded .71 after four observation occasions, .75 after five observations, and .86 after 10 observations. Several studies have provided predictive validity evidence for the categories used in the CASES (e.g., Coker, Medley, & Soar, 1980; Spaulding & Papageorgiou, 1972).

Procedures for STARS & CASES

Classroom interactions are recorded for every 5 to 10 seconds for a minimum of 8 minutes or 50 samples on each observation occasion. Each observation sample records one behavior category of a preselected student using CASES then followed by one category using STARS. Data are recorded by marking category numbers and letters (i.e., 25 categories for STARS and 19 categories for CASES) and making tallies. Each observation sample contains information on a specific student and simultaneous teacher-student interactions. Six or more students are selected for each observation occasion.

It appears that the training of observers is extensive when STARS and CASES are used in conjunction, taking approximately 3 to 4 weeks.

OCEPT-Teacher Observation Protocol (O-TOP) (Wainwright, Flick, & Morrell, 2003)

This instrument was developed primarily to assess teachers of science and mathematics course in higher education, based on the work of Pibum et al.(2000) as well as Lawrenz et al. (2002). It has a five point Likert like scale of rating ('Not observed' to 4) for 10 different aspects of classroom activities (e.g., This lesson encouraged students to seek and value various modes of investigation or problem solving). The instrument also allows to record 18 types of instruction used during the observed session such as lecture, problem modeling and small group discussion. In addition, this instrument offers a follow-up interview protocol (OCEPT Teacher Interview Protocol, OTIP) that validates the observational data and adds the instructor's perspective. However, this instrument is not able to provide quantifiable data on the frequency of various teacher and student behaviors and/or interactions. Percent agreement among observers reached 100% for 8 items while 71% and 57% for two items (one on metacognition and the other on student perceptions), respectively. It has content validity examined by experts in the science and mathematics (Morrell, Wainwright, Flick, 2004).

The Teacher Interview Protocol (OTIP) contains open-ended questions within the ten categories of the classroom observation protocol by addressing how the focal teacher's instruction supports student thinking, social skills and collaboration, and content knowledge.

Table 5 Summary of Recommended Classroom Observation Protocols

Measure	What is involved?	How long does it take to administer?	How much training is needed?	What are psychometric properties?	Strengths and weaknesses
Stallings Observation System	*Classroom Snapshot* records the frequency of different types of classroom behavior. *Five-Minute Interaction* captures information on each statement during observation in terms of (1) Who?, (2) To whom?, (3) What?, and (4) How? The Environmental checklist records arrangements and availability of instructional and resource materials.	*Classroom Snapshot* is completed five times a class period on at least two consecutive days per classroom. *Five-Minute Interaction:* same as the Classroom Snapshot	Training for the entire system took 7 days to yield reliability above .70 (Stallings, 1975). Revised Stallings Observation System took five days of training to yield reliability estimates above .86 (Knight, 2001). Classroom Snapshot requires two half day sessions for acceptable reliability (Knight, 2001). Five-Minute Interaction requires five days of training to achieve consistency among observers above 85%.	Reliability estimate for the original system is above .70 for each code with 7 days of training. The reliability coefficients for revised Stallings Observation System (Knight, 2001) with a computerized system were above .86 with a five-day training. The reliabilities are above .90 for Classroom Snapshot and .86 for Five-Minute Interaction (Knight, 2000).	<u>Strengths</u> It has been used extensively in many studies (K through 12th grades in almost all subjects). Sound reliability and validity evidence is available. It is available in different format: paper and pencil, scantron, computer version. <u>Weaknesses</u> The administration of the

				Validity evidence is available for school effectiveness (Ellett & Teddlie, 2003; Stallings, 1975)	entire system could be time-intensive.
Extra Teacher Project Observation Instrument	*Classroom Observation* records the frequency and type of classroom behavior on a surprise visit. *Pupil and Teacher Interviews* Record information from unobserved classroom behavior *Exercise book examination* assesses recent work and teacher feedback	60 minutes.	1 day training.	Reliability and validity information is not available yet.	<u>Strengths</u> Provides detailed information in a single visit. Avoids observer effect <u>Weaknesses</u> Validity and reliability not yet available.
Special Strategies Observation System – Revised (SSOS-R)	*Class Observation Schedule* is based on two well-established observation instruments: Stallings Observation System and Classroom Activity Record,	60 minutes	*Class Observation Schedule* requires 12 hour training. It appears that *QAIT* requires	For the QAIT instrument, reliability estimate for all 40 items was .96 (each of the four constituent scales ranged from .69 to .95).	<u>Strengths</u> This system was designed based on highly regarded previous observation systems.

	providing data on frequencies of different teaching behaviors. *QAIT* provides information on quality of instruction, which is based on Slavin's review of literature on effective teaching and learning practices. It also allows focus on a few target students. The classroom environment checklist records classroom environment and resources available.	extensive training for acceptable reliability. However, no specific information is published on this.	For the Classroom Observation Schedule reliability estimates are as follows: .51 for the activity code section coefficient, .76 for the student engagement, and .76 for the grouping section and .38 for the number of students section (Hughes et al., 2005). Several studies have demonstrated its concurrent and predictive validity evidence (Nesselrodt & Schaffer, 2000a, 2000b; Cowley et al., 2002).	Class Observation Schedule does not require extensive training. SSOS-R, QAIT in particular, provides data on quality of instruction. However, due to its high inference nature, more extensive training is required. <u>Weaknesses</u> Reliability estimates for a couple of subscales for the Classroom Observation Schedule are low. Thus, to achieve acceptable reliability, it may take extensive training.

Virgilio Teacher Behavior Inventory (VTBI)	It records three aspects of teacher effectiveness: classroom management, instruction, and classroom climate.	One class period	No published information on the intensity of observer training is available.	-Classroom management: .85 -Instruction .96 -Classroom climate: .85 Concurrent validity evidence was obtained from a positive correlation between VTBI and Stallings Classroom Snapshot (rs =.64, .63, .60, and .50 for total index, classroom management, instruction, and classroom climate	<u>Strengths</u> It is based on a review of research literature on teacher effectiveness. <u>Weaknesses</u> Due to a rating system, it is likely to require extensive training of observers. It provides information on teacher behaviors, but not on student behaviors.
Classroom Observations Keyed for Effectiveness Research (COKER, Coker & Coker, 1979)	It records teacher and student interactions, affective behaviors, and grouping configurations. Section A records teacher and student interaction types while Section B codes teacher and student affective behaviors.	The total observation time per visit is 10 minutes (two 5 minute observations): The actual time spent in the classroom could range from 20-25	Published information on the intensity of observer training is not readily available.	Published information on the intensity of observer training is not readily available.	<u>Weaknesses</u> Published information on statistical properties and intensity of the training is not readily available.

		minutes (Medley, Coker, & Soar, 1984).			
Spaulding Teacher Activity Recording Schedule (STARS, Spaulding, 1982) & Coping Analysis Schedule for Educational Settings (CASES, Simon & Boyer, 1967).	STARS provides information on the cognitive instructional methods, affective relationships, and behavior control strategies of teachers. 25 behavior patterns are included in the STARS. Time sampling technique is used (e.g., every 10 second), to yield frequency information on teacher and student response behaviors.	A minimum of 8 minutes or 50 samples on each observation occasion	It takes two to three weeks to yield a reliability estimate above .80.	STARS Reliability of observers ranged from .70s to .90s. Generalizability analysis yielded a correlation coefficient of .80 using four observations, and .90 using eight observations. Five of the 25 behavior categories showed convergent and discriminant validity.	<u>Strengths</u> STARS is developed based on comprehensive teacher-student interaction study (Spaulding, 1963).
	CASES records student classroom behaviors prespecified in 19 patterns. Time sampling technique can be	A minimum of 8 minutes or 50 samples on each observation occasion	It takes approximately 25 to 30 hours to yield a reliability estimate above .80.	CASES Four observation occasions yielded .71, five observations .75, and 10 observations .86.	CASES is developed based on sound theoretical background. When STARS and CASES

	used (e.g., every 10 second) to yield frequency information on student behaviors.			Several studies have provided predictive validity evidence for the categories used in the CASES .	are used in conjunction, they provide information on teacher and student behaviors. Weaknesses When both STARS and CASES are used in conjunction, they require extensive training (3-4 weeks). It takes several observations to yield acceptable reliability.
OCEPT-Teacher Observation Protocol (O-TOP)	It provides data on 10 different aspects of classroom activities and 18 instructional type.	One class period	No published information is available.	Percent agreement among observers reached 100% for 8 items while 71% and 57% for two items, respectively. It has content validity is noted (Morrell, Wainwright, Flick, 2004), but other concurrent and predictive	Strengths The teacher interview can add depth to the observation data. Weaknesses It is primarily designed to

		validity information is not available.	evaluate science and mathematics teachers in higher education. It is high inference system such that it appears to require extensive training for acceptable reliability.

RECOMMENDATIONS

The principle outcomes measure for the World Bank multi-country impact study should be a quantitative low-inference measure providing comparable data appropriate for each participating country. The Stallings Observation System meets these criteria and has good reliability and validity. There are two main weaknesses with this measure – cost and applicability to developing countries. Suggested modifications to the measure are offered in the following section.

Adapting Stallings Classroom Observation System

1) Reducing time taken to conduct the measure

The administration of the entire Stallings Classroom Observation System is costly and time-intensive. Of the three components, the Environmental Checklist is relatively quick to administer. Reducing the length of the checklist to key components of the classroom environment can reduce its length further (see accompany documents). With the remaining two instruments, there is a challenge in implementation. According to the published recommendation, it takes four days of observation for a classroom, two consecutive days using the Snapshot and two consecutive days using the Five-Minute Interaction. Committing four days for a classroom may not be practical in a large scale study. There are also disadvantages in terms of teachers adapting their behavior whilst being observed (see below). There are two options for reducing the amount of time spent on the observation measures.

a) Observations can be conducted on one day only.

We were unable to find any published information about the reliability of conducting the measure on one day, rather than on two days. Before this measure is adopted, pilot work should establish the reliability of a reduced measure. The reliability of the measure, and the amount of information captured, may be improved by including adaptations from the Extra Teacher Project Observation System. Specifically, the retrospective collection of data through the pupil and teacher interviews and through the examination of textbooks will add useful information to the measure. Reliability data on these measures will soon be available.

In addition, Five Minute Interaction measure could be dropped from the battery if necessary. The Stallings Snapshot may capture sufficient information to be used as the sole observation measure and offers many strengths: (1) it provides quantitative information on classroom interactions that has been proven to be reliable in many contexts; (2) it is easy to use; (3) it requires just two half-day training sessions to yield reliable data (Knight, 2000); (4) the categories in the Snapshot can be easily adapted to fit the goals of a study.

b) Adapt measures to the developing country context.

The quantity of actual instructional time is a stronger determinant of achievement in developing countries compared to developed countries. The Stallings Observation System could be adapted to assess this by (1) directly assessing the amount of time spent in instructional and non-instructional activity in a given class and (2) using retrospective data collection methods from the Extra Teacher Project (pupil and student interviews, exercise book examination) to collect data that are less influenced by the presence of an observer. In addition, it may be informative to include a category such as 'Noninstructional Activity' in the observation schedule to capture the extent to which teacher are engaged in activities that are not related to instruction.

c) Establish cross-country consistency

Where the assessment goal is to compare the classroom behavior directly between two or more countries (rather than, for example, comparing the impact of an intervention on classroom behavior across countries), it is imperative that behavior is coded in the same way across all contexts. There are a number of ways in which this can be achieved

(i) Training instructions could be piloted in a number of countries and modifications made to a master version of the training procedure which is then translated (but not adapted) for use in participating countries

(ii) Reliability of coding between countries could be formally assessed by observers working in more than one country

(iii) If classroom behavior categories are aggregate into 'mega-variables' (Knight, 2000), inter-rater reliability is likely to be higher. This is because behaviors may be coded into slightly different sub-categories by two raters, but are less likely to be wrongly coded into a very different higher level category.

NEXT STEPS

(1) Categories of behavior in the Stallings Snapshot should be modified in accordance with the aims of the project. The Snapshot is designed to be flexible in this way. One such modification includes recording Nonintructional Activity, but others may emerge from clearly defined aims of the project (see Appendix A). Behavioral categories chosen should be reviewed by educational professionals in host countries to ensure they are meaningful in participating countries.

(2) Pilot work in target countries should establish the reliability of conducting a single Stallings Snapshot, rather than conducting it on two consecutive days. This pilot should also assess the amount of information lost by omitting the repeat measure.

(3) Further pilot work should develop standard training procedures that are suitable for all participating countries and establish the cross-country reliability of the observation measure.

(4) If results from the Extra Teacher Project show that retrospective assessments of classroom behavior are valid and reliable, these should be included in the assessment battery.

References – Reading Assessments

Abadzi, H., Crouch, L., Echegaray, M., Pasco, C., & Sampe, J. (2005). Monitoring basic skills acquisition through rapod learning assessments: A case study from Peru. *Prospects, XXXV*(2), 137-156.

Abu-Rabia, S. (1997). The need for cross-cultural considerations in reading theory: the effects of Arabic sentence context in skilled and poor readers. *Journal of Research in Reading, 20*(2), 137-147.

Abu-Rabia, S. (1997). Reading in Arabic orthography: The effect of vowels and context on reading accuracy of poor and skilled native Arabic readers. *Reading and Writing: An Interdisciplinary Journal, 9*, 65-78.

Abu-Rabia, S. (1999). The effect of Arabic vowels on the reading comprehension of second- and sixth-grade native Arab children. *Journal of Psycholinguistic Research, 28*(1), 93-101.

Abu-Rabia, S. (2001). The role of vowels in reading Semitic scripts: Data from Arabic and Hebrew. *Reading and Writing: An Interdisciplinary Journal, 14*, 39-59.

Abu-Rabia, S. (2002). Reading a root-based-morphology language: the case of Arabic. *Journal of Research in Reading, 25*(3), 299-309.

Abu-Rabia, S., & Siegel, L. S. (2002). Reading, syntactic, orthographic, and working memory skills of bilingual Arabic-English speaking Canadian children. *Journal of Psycholinguistic Research, 31*, 661-678.

Adams, M. J. (2004). Modeling the connections between word recognition and reading. In R. B. Ruddell & N. J. Unrau (Eds.), *Theoretical models and processes of reading* (Fifth ed., pp. 1219-1243). Newark, DE: International Reading Association.

Adroin, S. P., Witt, J. C., Suldo, S. M., Connell, J. E., Koenig, J. L., Resetar, J. L., et al. (2004). Examining the incremental benefits of administering a maze and three versus one curriculum-based measurement reading probes when conducting universal screening. *School Psychology Review, 33*(2), 218-233.

Aro, M., & Wimmer, H. (2003). Learning to read: English in comparison to six more regular orthographies. *Applied Psycholinguistics, 24*, 621-635.

Barnea, A., & Breznitz, Z. (1998). Phonological and orthographic processing of Hebrew words: Electrophysiological aspects. *The Journal of Genetic Psychology, 159*(4), 492-504.

Bear, D. R. (2001). "Learning to fasten the seat of my union suit without looking around": The synchrony of literacy development. *Theory Into Practice, XXX*(3), 149-157.

Berninger, V. W., Abbott, R. D., Billingsley, F., & Nagy, W. (2001). Processes underlying timing and fluency of reading: Efficiency, automaticity, coordination, and morphological awareness. In M. Wolf (Ed.), *Dyslexia, fluency, and the brain* (pp. 383-413). Maryland: York Press.

Bowers, P. G., & E., N.-C. (2002). The role of naming speed within a model of reading acquisition. *Reading and Writing: An Interdisciplinary Journal, 15*, 109-126.

Breznitz, Z. (1997). Effects of accelerated reading rate on memory for text among dyslexic readers. *Journal of Educational Psychology, 89*(2), 289-297.

Breznitz, Z. (1997). Enhancing the reading of dyslexic children by reading acceleration and auditory masking. *Journal of Educational Psychology, 89*(1), 103-113.

Breznitz, Z. (1997). Reading rate acceleration: Developmental aspects. *The Journal of Genetic Psychology, 158*(4), 427-441.

Breznitz, Z. (2001). The determinants of reading fluency: A comparison of dyslexic and average readers. In M. Wolf (Ed.), *Dyslexia, fluency, and the brain* (pp. 245-276). Maryland: York Press.

Breznitz, Z. (2002). Asynchrony of visual-orthographic and auditory-phonological word recognition processes: An underlying factor in dyslexia. *Reading and Writing: An Interdisciplinary Journal, 15*, 15-42.

Breznitz, Z., & Berman, L. (2003). The underlying factors of word reading rate. *Educational Psychology Review, 15*(3), 247-265.

Breznitz, Z., & Share, D. (2002). Introduction on timing and phonology. *Reading and Writing: An Interdisciplinary Journal, 15*, 1-3.

Buly, M. R., & Valencia, S. (2003). *Meeting the needs of failing readers: Cautions and considerations for state policy*: Center for the Study of Teaching and Policy: A National Research Consortium.

Chall, J. S. (1996). *Learning to read: The great debate* (Third ed.). New York: Harcourt Brace College Publishers.

Clay, M. (1967). The reading behaviour of five year old children: A research report. *New Zealand Journal of Educational Studies, 2*(1), 11-31.

Clay, M. (1991). *Becoming literate: The construction of inner control.* Auckland, NZ: Heinemann.

Cunningham, A. E., & Stanovich, K. E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33*(6), 934-945.

DeFord, D. E. (1991). Fluency in initial reading instruction: A reading recovery lesson. *Theory Into Practice, XXX*(3), 201-210.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 2*, 219-232.

Durgunoglu, A. Y., & **Öney, B.** (1999). A cross-linguistic comparison of phonological awareness and word recognition. *Reading and Writing: An Interdisciplinary Journal, 11*, 281-299.

Ehri, L. C. (1979). Linguistic insight: Threshold of reading acquisition. In T. G. Waller & G. E. MacKinnon (Eds.), *Reading research: Advances in theory and practice* (Vol. 1, pp. 103-148). New York: Academic.

Ehri, L. C. (1995). Phases of development in learning to read words by sight. *Journal of Research in Reading, 18*(2), 116-125.

Ellis, N. C., & Hooper, A. M. (2001). Why learning to read is easier in Welsh than in English: Orthographic transparency effects evinced with frequency-matched tests. *Applied Psycholinguistics, 22*, 571-599.

Ferreiro, E., & Teberosky, A. (1982). *Literacy before schooling* (K. G. Castro, Trans.). New Hampshire: Heinemann Educational Books.

Foulin, J. N. (2005). Why is letter-name knowledge such a good predictor of learning to read? *Reading and Writing, 18*, 129-155.

Frost, R. (1994). Prelexical and postlexical strategies in reading: Evidence from a deep and a shallow orthography. *Journal of Experimental Psychology, 20*(1), 116-129.

Fuchs, L. S., & Fuchs, D. (1999). Monitoring student progress toward the development of reading competence: A review of three forms of classroom-based assessment. *School Psychology Review, 28*(4), 659-671.

Fuchs, L. S., Fuchs, D., Hosp, M. K., & Jenkins, J. R. (2001). Oral reading fluency as an indicator of reading competence: A theoretical, empirical, and historical analysis. *Scientific Studies of Reading, 5*(3), 239-256.

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. *Remedial and Special Education, 9*(2), 20-28.

Geva, E., & Siegel, L. S. (2000). Orthographic and cognitive factors in the concurrent development of basic reading skills in two languages. *Reading and Writing: An Interdisciplinary Journal, 12*, 1-30.

Gillingham, M. G., & Garner, R. (1992). Reader's comprehension of mazes embedded in expository text. *Journal of Educational Research, 85*(4), 234-241.

Good, R. H., & Kaminski, R. A. (Eds.). (2002). *Dynamic Indicators of Basic Early Literacy Skills (DIBELS, 6th ed.).* Eugene, OR: Institute for the Development of Educational Achievement.

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5*(3), 257-288.

Goswami, U., & Bryant, P. E. (1990). *Phonological skills and learning to read*. Hillsdale, NJ: Lawrence Erlbaum.

Goswami, U., Ziegler, J. C., Dalton, L., & Schneider, W. (2003). Nonword reading across orthographies: How flexible is the choice of reading units? *Applied Psycholinguistics, 24*, 235-247.

Gough, P. B. (1995). The new literacy: caveat emptor. *Journal of Research in Reading, 18*(2), 79-86.

Gupta, A. (2004). Reading difficulties of Hindi-speaking children with developmental dyslexia. *Reading and Writing: An Interdisciplinary Journal, 17*, 79-99.

Hamilton, C., & Shinn, M. R. (2003). Characteristics of word callers: An investigation of the accuracy of teachers' judgment of reading comprehension and oral reading skills. *School Psychology Review, 32*(2), 228-240.

Hasbrouck, J., & Tindal, G. A. (2006). Oral reading fluency norms: A valuable assessment tool for reading teachers. *International Reading Association, 59*(7), 636-644.

Hecht, S. A., Burgess, S. R., Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2000). Explaining social class differences in growth of reading skills from beginning kindergarten through fourth grade: The role of phonological awareness, rate of access, and print knowledge. *Reading and Writing: An Interdisciplinary Journal, 12*, 99-127.

Hiebert, E. H. (2002). Standards, assessments, and text difficulty. In A.E.Farstrup & J.Samuels (Eds.), *What research has to say about reading instruction* (Third ed., pp. 337-369). Delaware: International Reading Association.

Hintze, J. M., Callahan III, J. E., Matthews, W. J., Williams, S. A. S., & Tobin, K. G. (2002). Oral reading fluency and prediction of reading comprehension in African American and Caucasian elementary school children. *School Psychology Review, 31*(4), 540-553.

Holopainen, L., Ahonen, T., & Lyytinen, H. (2001). Predicting delay in reading achievement in a highly transparent language. *Journal of Learning Disabilities, 34*(5), 401-413.

Hosp, M. K., & Fuchs, L. S. (2005). Using CBM as an indicator of decoding, word reading, and comprehension: Do the relations change with grade? *School Psychology Review, 34*(1), 9-26.

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Accuracy and fluency in list and context reading of skilled and RD groups: Absolute and relative performance levels. *Learning Disabilities Research and Practice, 18*(4), 237-245.

Jenkins, J. R., Fuchs, L. S., van den Broek, P., Espin, C., & Deno, S. L. (2003). Sources of individual differences in reading comprehension and reading fluency. *Journal of Educational Psychology, 95*(4), 719-729.

Jenkins, J. R., Zumeta, R., & Dupree, O. (2005). Measuring gains in reading ability with passage reading fluency. *Learning Disabilities Research and Practice, 20*(4), 245-253.

Juel, C. (1988). Learning to read and write: A longitudinal study of 54 children from first through fourth grades. *Journal of Educational Psychology, 80*(4), 437-447.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329-354.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*(1), 122-149.

Just, M. A., Carpenter, P. A., & Keller, T. A. (1996). The capacity theory of comprehension: New frontiers of evidence and arguments. *Psychological Review, 103*(4), 773-780.

Kame'enui, E. J., & Simmons, D. C. (2001). Introduction to this special issue: The DNA of reading fluency. *Scientific Studies of Reading, 5*(3), 203-210.

Karanth, P., Mathew, A., & Kurien, P. (2004). Orthography and reading speed: Data from native readers of Kannada. *Reading and Writing: An Interdisciplinary Journal, 17*, 101-120.

Katzir, T., Kim, Y.-S., Wolf, M., Kennedy, B., Lovett, M., & Morris, R. (2006). The relationship of spelling recognition, RAN, and phonological awareness to reading skills in older poor readers and younger reading-matched controls. *Reading and Writing, DOI 10.1007/s11145-006-9013-2*.

Katzir, T., Shaul, S., Breznitz, Z., & Wolf, M. (2004). The Universal and the unique in dyslexia: A cross-linguistic investigation of reading and reading fluency in Hebrew- and English-speaking children with reading disorders. *Reading and Writing: An Interdisciplinary Journal, 17*, 739-768.

Kintsch, E. (2005). Comprehension theory as a guide for the design of thoughtful questions. *Top Language Disorders, 25*(1), 51-64.

Kintsch, W. (1986). Learning from text. *Cognition and Instruction, 3*(2), 87-108.

Kotoulas, V. (2004). The development of phonological awareness throughout the school years: The case of a transparent orthography. *Educational Studies in Language and Literature, 4*, 183-201.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology, 6*, 293-323.

Liow, S. J. R., & Lee, L. C. (2004). Metalinguistic awareness and semi-syllabic scripts: Children's spelling errors in Malay. *Reading and Writing: An Interdisciplinary Journal, 17*, 7-26.

Logan, G. D. (1997). Automaticity and reading: Perspectives from the instance theory of automatization. *Reading and Writing Quarterly, 13*(2), 1057-3569.

MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: Comment on Just and Carpenter (1992) and Waters and Caplan (1996). *Psychological Review, 109*(1), 35-54.

Näslund, J. C., & Schneider, W. (1996). Kindergarten letter knowledge, phonological skills, and memory processes: Relative effects on early literacy. *Journal of Experimental Child Psychology, 62*, 30-59.

Oney, B., & Durgunoglu, A. Y. (1997). Beginning to read in Turkish: A phonologically transparent orthography. *Applied Psycholinguistics, 18*, 1-15.

Parker, R., Hasbrouck, J. E., & Tindal, G. (1992). The maze as a classroom-based reding measure: Construction methods, reliability, and validity. *The Journal of Special Education, 26*(2), 195-218.

Perfetti, C. A. (1977). Language comprehension and fast decoding: Some psycholinguistic prerequisites for skilled reading comprehension. In J.T.Guthrie (Ed.), *Cognition, curriculum, and comprehension* (pp. 20-41). Newark, DE: International Reading Asoociation.

Perfetti, C. A. (1985). *Reading ability*. London: Oxford.

Perfetti, C. A. (1988). Verbal efficiency in reading ability. In M. Daneman, G.E.Mackinnon & T.G.Waller (Eds.), *Reading research: Advances in theory and practice*. Boston, MA: Acadmeic Press.

Pratham (2005). Annual status of education report (ASER) Retrieved 18 August, 2006, 2006, from http://www.pratham.org/aserrep.php

Progress, N. A. o. E. (2005). *Fourth-grade students reading aloud: NAEP 2002 special study of oral reading*. Washington, D.C.: National Center for Education Statistics.

Rupley, W. H., Willson, V. L., & Nichols, W. D. (1998). Exploration of the developmental components contributing to elementary school children's reading comprehension. *Scientific Studies of Reading, 2*(2), 143-158.

Saiegh-Haddad, E. (2003). Bilingual oral reading fluency and reading comprehension: The case of Arabic/Hebrew (L1)-English (L2) readers. *Reading and Writing: An Interdisciplinary Journal, 16*, 717-736.

Saiegh-Haddad, E. (2005). Correlates of reading fluency in Arabic: Diglossic and orthographic factors. *Reading and Writing, 18*(559-582).

Samuels, S. J. (2002). Reading fluency: Its development and assessment. In A.E.Farstrup & J.Samuels (Eds.), *What research has to say about reading instruction* (Third ed., pp. 166-183). Delaware: International Reading Association.

Samuels, S. J. (2004). Toward a theory of automatic information processing in reading, revisited. In R. B. Ruddell & N. J. Unrau (Eds.), *Theoretical models and*

processes of reading (Fifth ed., pp. 1127-1148). Newark, DE: International Reading Association.

Scarborough, H. S. (1998). Early identification of children at risk for reading disabilities: Phonological awareness and some other promising predictors. In B. K. Shapiro, P. J. Accardo & A. J. Capute (Eds.), *Specific reading disability: A view of the spectrum* (pp. 77-121). Timonium, MD: York Press.

Sénéchal, M., & LeFevre, J. (2001). Storybook reading and parent teaching: Links to language and literacy development. In P. R. Britto & J. Brooks-Gunn (Eds.), *The roles of family literacy environments in promoting young children's emergent literacies* (pp. 39-52). San Francisco: Jossey-Bass.

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*, 143-174.

Share, D. L., Jorm, A. F., Maclean, R., & Matthews, R. (1984). Sources of individual differences in reading acquisition. *Journal of Educational Psychology, 76*(6), 1309-1324.

Shimron, J., & Sivan, T. (1994). Reading proficieny and orthography: Evidence from Hebrew and English. *Language Learning, 44*, 5-27.

Sprugevica, I., & Hoien, T. (2003). Enabling skills in early reading acquisition: A study of children in Latvian kindergartens. *Reading and Writing: An Interdisciplinary Journal, 16*, 159-177.

Stanovich, K. E. (1991). Word recognition: Changing perspectives. In R. Barr, M. L. Kamil, P. Mosenthal & P. D. Pearson (Eds.), *Handbook of reading research* (Vol. 2, pp. 418-452). New York: Longman.

Stanovich, K. E., & Stanovich, P. J. (1995). How research might inform the debate about early reading acquisition. *Journal of Research in Reading, 18*(2), 87-105.

Stuart, M. (1995). Through printed words to meaning: issues of transparency. *Journal of Research in Reading, 18*(2), 126-131.

Tan, A., & Nicholson, T. (1997). Flashcards revisited: Training poor readers to read words faster improves their comprehension of text. *Journal of Educational Psychology, 89*(2), 276-288.

Taouk, M., & Colthear, M. (2004). The cognitive processes involved in learning to read in Arabic. *Reading and Writing: An Interdisciplinary Journal, 17*, 27-57.

Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of Word Reading Efficiency (TOWRE)*. Austin, TX: Pro-Ed.

Tressoldi, P. E. (personal communication, August 27, 2006).

Tressoldi, P. E., Stella, G., & Faggella, M. (2001). The development of reading speed in Italians with dyslexia: A longitudinal study. *Journal of Learning Disabilities, 34*(5), 414-417.

Vaid, J., & Padakannaya, P. (2004). Introduction. *Reading and Writing: An Interdisciplinary Journal, 17*, 1-6.

Valencia, S. W., & Buly, M. R. (2005). Behind test scores: What struggling readers really need. In S. J. Barrentine & S. M. Stokes (Eds.), *Reading assessment: Principles and practices for elementary teachers* (Second ed., pp. 134-146). Delaware: International Reading Association.

Vasanta, D. (2004). Processing phonological information in a semi-syllabic script: Developmental data from Telugu. *Reading and Writing: An Interdisciplinary Journal, 17*, 59-78.

Verhoeven, L. (1998). A future perspective on literacy in Europe. *Peabody Journal of Education, 73*(3&4), 127-144.

Wagner, D. A. (1993). *Literacy, culture, and development: Becoming literate in Morocco*. New York: Cambridge University Press.

Waters, G. S., & Caplan, D. (1996). The capacity theory of sentence comprehension: Critique of Just and Carpenter (1992). *Psychological Review, 103*(4), 761-772.

Waters, G. S., & Caplan, D. (1996). The measurement of verbal working memory capacity and its relation to reading acquisition. *The Quarterly Journal of Experimental Psychology, 49A*(1), 51-79.

Wesseling, R., & Reitsma, P. (2000). The transient role of explicit phonological recoding for reading acquisition. *Reading and Writing: An Interdisciplinary Journal, 13*, 313-336.

West, R. F., Stanovich, K. E., Feeman, D. J., & Cunningham, A. E. (1983). The effect of sentence context on word recognition in second- and sixth-grade children. *Reading Research Quarterly, 19*(1), 6-15.

Whitney, P., Arnett, P. A., Driver, A., & Budd, D. (2001). Measuring central executive functioning: What's in a reading span? *Brain and Cognition, 45*, 1-14.

Wiley, H. I., & Deno, S. L. (2005). Oral reading and maze measures as predictors of success for English learners on a state standards assessment. *Remedial and Special Education, 26*(4), 207-214.

Williams, K. T. (2004). *Reading Fluency Indicator*. Minnesota: AGS Publishing.

Wimmer, H., & Goswami, U. (1994). The influence of orthographic consistency on reading development: word recognition in English and German children. *Cognition, 51*, 91-103.

Wimmer, H., & Mayringer, H. (2001). Is the reading-rate problem of German dyslexic children caused by slow visual processes? In M. Wolf (Ed.), *Dyslexia, fluency, and the brain* (pp. 93-102). Maryland: York Press.

Wimmer, H., Mayringer, H., & Landerl, K. (1998). Poor reading: A deficit in skill-automatization or a phonological deficit? *Scientific Studies of Reading, 2*(4), 321-340.

Wolf, M., Bowers, P. G., & Biddle, K. (2000). Naming-speed processes, timing, and reading: A conceptual review. *Journal of Learning Disabilities, 33*, 387-407.

Wolf, M., & Katzir-Cohen, T. (2001). Reading fluency and its intervention. *Scientific Studies of Reading, 5*(3), 211-239.

Yovanoff, P., Duesbery, L., Alonzo, J., & Tindal, G. (2005). Grade-level invariance of a theoretical causal structure predicting reading comprehension with vocabulary and oral reading fluency. *Educational Measurement: Issues and Practice, 24*, 4-12.

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin, 131*(1), 3-29.

References – Classroom Observation Review

Ali, M. and T. Reed (1994) A School and Parental Survey of Book Provision Issues in NWFP, International Book Development, Ltd.

Anderson, L., Ryan, D., & Shapiro, B. (1989). *The IEA classroom environment study*. New York: Pergamon.

Baker, V. (1988) "Schooling and disadvantage in Sri Lankan and other rural situations," Comparative Education, 24, 3, Pp. 377-388.

Avot, A. (2004). Factors affecting actual instructional time in primary schools: A literature review. Report for the World Bank-IBE study on instructional time (ED871-138-3).

Berliner, D., & Biddle, B. (1995). Tempus educare. In P. Peterson & H. Walberg IEDS.), *Research on teaching: Concepts, findings, and implications* (pp. 769-818). Berkeley, CA: McCutchan.

Boehm, A. E. & Weinberg, R. A. (1997). *The classroom observer: Developing observation skills in early childhood settings*. New York, Teachers College Press.

Bramlett, R. K., & Barnett, D. W. (1993). The development of a direct observation code for use in preschool settings. *School Psychology Review, 22*, 49-62.

Brophy, J., & Good, T (1986). Teacher behavior and student achievement. In M. C. Witrock (Ed.), *Handbook of research on teaching* (3rd ed.). New York: Macmillan.

Carroll, J. B. (1963). A model of school learning. *Teachers College Record, 64*, 723-733.

Dia, E. C. (2003). Instructional time in primary school: The cases of Burkina Faso and the Gambia. Paper prepared for HDNED. Washington DC: World Bank.

Ellett, C. D., & Teddlie, C. (2003). Teacher evaluation, teacher effectiveness and school effectiveness: Perspectives from the USA. *Journal of Personnel Evaluation in Education, 17* (1), 101-128.

Evertson, C., & Burry, J. (1989). Capturing classroom context: The observation system as lens for assessment. *Journal of Personnel Evaluation in Education, 2,* 297-320.

Fuller, B., Dellagnelo, L., Strath, A., Bastos, E. S. B., Maia, M. H., de Matoes, K. S. L., Portela, A. L., Vieira, S. L. C. (1999). How to raise children's early literacy? The influence of family, teacher, and classroom in northeast Brazil. Comparative Education Review, 43, 1, 1-35.

Coker, H., Medley, D. M., & Soar, R. S. (1980). How valid are expert opinions about effective teaching? *Phi Delta Kappan, 62,* 131-134, 149.

Cowley, K. S., Meehan, M. L., Finch, N., Chadwick, K., Howley, C., Riffle, J., et al.
(2002). *Comprehensive evaluation of the Kentucky extended school services program.*
Charleston, WV: AEL.

Good, T. L., & Brophy, J. E. (1994). *Looking in classrooms* (6th ed.). New York: Harper-Collins.

Good, T. L., & Brophy, J. E. (2000). Motivation. In T. Good & J. Brophy, Eds., Looking in classrooms (8th ed.), 217-267. New York, NY: Longman

Goodlad, J. (1980). *A place called school*. New York: McGraw-Hill.

Heyneman, S. P., & Loxley, W. A. (1983). The effect of primary-school quality on academic achievement across twenty-nine high- and low income countries. *American Journal of Sociology, 88* (6), 1162-1194.

Hillberg, R. S., Waxman, H. C., & Tharp, R. G. (2004). Introduction: Purposes and perspectives on classroom observation research. In H. C. Waxman, R. G. Tharp, & R. S.

Hilberg (Eds.), *Observational research in U.S. classrooms: New approaches for understanding cultural and linguistic diversity* (pp. 1-20). Cambridge, UK: Cambridge University Press.

Hughes, G. K., Cowlet, K. S., Copley, L. D., Finch, N. L., Meehan, M. L. (2005). Evaluation of a multi-school pilot project designed to close achievement gaps. A paper presented at 2005 National Evaluation Institute, Memphis, TN.

Knight, S. L. (April 2001). Using technology to update traditional classroom observation instruments. A paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Linden, T. (2001) Double-Shifts Secondary School: Possibilities and Issues. Secondary Education
Series, Washington D.C.: World Bank.

Medley, D. M. (1982). Systemic observation. In H. E. Mitzel, J. H. Best, & W. Rabinowitz (Eds.), *Encyclopedia of educational research* (5th ed., Vol. 4, pp. 1841-1851). New York: The Free Press.

Medley, D. M., Coker, H., & Soar, R. S. (1984). *Measurement-based evaluation of teacher performance: An empirical approach*. New York: Longman Inc.

Meehan, M. L., Cowley, K. S., Finch, N. L., Chadwick, K. L., Ermolov, L. D., & Riffle, M. J. S. (2004). Special Strategies Observation System-Revised: A useful tool for educational research and evaluation. A paper presented at the American Educational Research Association, Montreal, 2005. Eric document : ED484936.

Morrell, P. D., Wainwright, C., Flick, L. (2004). Reform teaching strategies used by student teachers. *School Science and Mathematics, 104* (5), 199-213.

Nesselrodt, P. S., & Schaffer, E. C. (1993, April). *The ISERP programme: A revised classroom observation instrument.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta, GA.

Nesselrodt, P. S., & Schaffer, E. C. (2000a). *External evaluation of Kentucky's extended school services, spring 2000: Phase 1 – final report.* Carlisle, PA: Author.

Nesselrodt, P. S, & Schaffer, E. C. (2000b). *External evaluation of Kentucky's extended*
school services, spring 2000: Phase 1 – final report, part 2. Carlisle, PA: Dickinson
College.

Pitkoff, E. (1993) Teacher Absenteeism: What Administrators Can Do, *NASSP Bulletin*, 77, 551, pp. 39-45.

PROBE - Public Report On Basic Education in India (1999). New Delhi: Oxford University Press.

Reimers, F. (1993). Time and Opportunity to Learn in Pakistan's Schools: Some Lessons on the Links between Research and Policy, *Comparative Education, 29, 2*, pp. 201-12.

Reynolds, A. (1992). What is competent beginning teaching? *A review of the literature. Review of Educational Research, 62,* 1-35.

Rogers, H. (2003) Indonesia Education Absenteeism Survey, World Bank and SMERU.

Schaffer, E., Nesselrodt, P., & Stringfield, S. (1991). *The groundings of an observation instrument: The teacher behavior-student learning research base of the Special Strategies Observation System.* Paper presented at the International School Effects Research Workshop, Kaohsiung, Taiwan.

Simon, A., & Boyer, E. G. (1967). *Mirrors for behavior: An anthropology of observation instruments* (Vol. 5). Philadelphia: Research for Better Schools, 1967.

Slavin, R. (1987). A theory of school and classroom organization. *Education Psychologist, 22,* 89-108.

Slavin, R. (1989). A theory of school and classroom organization. In R. Slavin (Ed.), *School and classroom organization.* Hillsdale, NJ: Erlbaum.

Smith, B. (2000). Quantity matters: Annual instructional time in an urban school system. *Educational Administration Quarterly, 36* (5), 652-682.

Spaulding, R. L. (1982). Gerneralizability of teacher behavior: Stability of observational data within and across facets of classroom environments. *Journal of Educational Research, 76*, 5-13.

Spaulding, R. L., & Papageorgiou, M. R. (1972). Effects of early intervention in the lives of disadvantaged children. ERIC Document ED 066 246.

Stallings, J. A. (1978). *Learning to look: A handbook on classroom observation and teaching models*. Belmont, CA: Wadsworth Publishing Company.

Stallings, J. (1980). Allocated academic learning time revisited, or beyond time on task. *Educational Researcher, 9* (11), 11-16.

Stallings, J., & Kaskowitz, D. (1974). *Follow through classroom observation evaluation,* 1972-1973 (SRI Project URU-7370). Menlo Park, CA: Stanford Research Institute.

Stallings, J., & Freiberg, 1. (1991). Observation for the improvement of teaching. In H. Waxman & H. Walberg (Eds.), *Effective teaching: Current research* (pp. 107-133). Berkeley: McCutchan.

Stallings, J. A., & Mohlman, G. G. (1988). Classroom observation techniques. In J. P. Keeves (Ed.), *Educational research, methodology, and measurement: An international handbook* (pp. 469-474). Oxford: Pergamon Press.

Stallings, J., Needels, M., & Stayrook, N. (1979). *How to Change the Process of Teaching Basic Reading Skills in Secondary Schools.* Final Report to the National Institute of Education. Menlo Park, CA: SRI International.

Stringfield, S., Winfield, L., Millsap, M. A., Puma, M. J., Gamse, B., & Randall, B. (1994). *Urban and suburban/rural special strategies for educating disadvantaged children: First year report.* Washington, DC: U. S. Department of Education.

Tan, J.P., Lane, J. and P. Coustère (1997) “Putting Inputs to Work in Elementary Schools: What Can Be Done in the Philippines”, Economic Development and Cultural Change, 45, 4, Pp. 857-879.

Teddlie, C., Kirby, P., & Stringfield, S. (1989). Effective versus ineffective schools: Observable differences in the classroom. *American Journal of Education, 97* (3), 221-236.

Teddlie, C., Virgiio, I., & Oescher, J. (1990). Development and Validation of the Virgilio Teacher Behavior Instrument. *Educational and Psychological Measurement, 50* (2), 421-430.

Tollefson, N., Lee, S. & Webber, L. (2001). *The consistency of systematic classroom observations in urban schools.* Kansas City, MO: Ewing Marion Kauffman Foundation. (ERIC Document Reproduction Service. No. ED 457 155).

Verwimp, P. (1999) "Measuring the Quality of Education at Two Levels: A Case Study of Primary Schools in Rural Ethiopia", International Review of Education, 45, 2, Pp. 167-196.

Vergilio, I. (1987). *An examination of the relationships among school effectiveness in elementary and junior high schools*. Doctoral dissertation, University of New Orleans.

Wainwright, C., Flick, L., Morrell, P. D. (2003). The development of instruments for assessment of instructional practices in standards based teaching. *The Journal of Mathematics and Science: Collaborative Explorations, 6*, 1-9.

Wainwright, C., Flick, L., Morrell, P. D., Schepige, A. (2004). Observation of reform teaching in undergraduate level mathematics and science courses. *School Science and Mathematics, 104*(7), 322-335.

Waxman, H. C., & Padron, Y. N. (1995). Improving the Quality of Classroom Instruction for Students at Risk of Failure in Urban Schools. *Peabody Journal of Education, 70* (2), 44-65.

Waxman, H. C., & Walberg, H. J. (Eds. 1991). *Effective Teaching: Current Research*. Berkeley, CA.: McCutchan Publishing.

Weinrott, M. R., Jones, R. R., & Boler, G. R. (1981). Convergent and discriminate validity of five classroom observation system: A secondary analysis. *Journal of Educational Psychology, 73*, 671-680.

World Bank (2001). Expanding and Improving Upper Primary Education in India, Washington, DC: World Bank.

***Appendix A*:** Assessing Early Reading Ability in Swahili and English: The Role of Letter Reading Fluency (LRF) and Oral Reading Fluency (ORF) and other Measures

Aim:

To develop and evaluate outcome measures for use in a multi-country study of Grade 2 reading ability.

Three different types of tests were assessed. Objectives are outlines for each set of tests

1) Letter Reading Fluency and Oral Reading Fluency

The present pilot study examines the use of Letter Reading Fluency (LRF) and Oral Reading Fluency (ORF) as valid measures of primary grade children's reading ability. Specifically the study addresses:

1. What is the performance of second and third grade children on Swahili and English LRF and ORF assessments?
2. For Swahili, what is the association of LRF and ORF with existing tests of reading ability in the Kenyan context? (I. e. is Oral reading fluency a valid measure of comprehension in languages with a shallow orthographies?)
3. Can LRF and ORF provide sensitive assessments for children with poor reading ability?
4. Does the addition of non-word reading add improve the validity of this assessment?

2) Written Group-Administered Tests of Reading Ability

1. What is the performance of second and third graders on the written reading tests in Swahili?
2. Does this test provide a sensitive and valid measure of reading ability.

3) Maze measure

1. What is the performance of second and third graders on the Maze test in Swahili and English?
2. Does this test provide a sensitive measure of reading ability in English and Swahili.
3. Does this test provide a more direct (and therefore more valid) measure of reading comprehension in English and Swahili?

Participants:

35 second graders and 86 third graders from 4 schools in Butere-Mumias District, Western Kenya. Although the aim was to develop tests appropriate for use in Grade 3, reading level in Grade 2 was found to be lower than expected in initial piloting. Grade 3 students were included to ensure assess psychometric properties of the tests at the upper end of the ability scale.

Assessing early reading ability in Swahili: Descriptive Statistics

The descriptive statistics for the Swahili reading assessments are presented in Table A1.

Swahili Letter Reading Fluency:

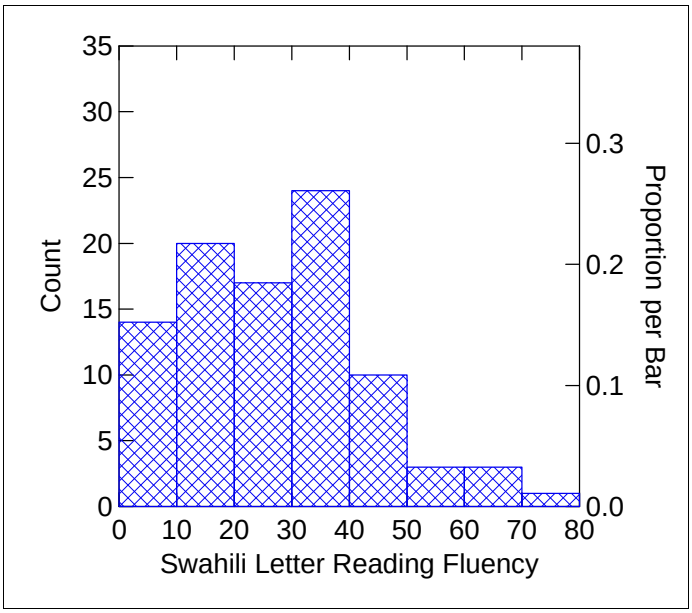
The Swahili letter reading fluency score indicates the speed and accuracy with which children read aloud randomly arranged letters of the Swahili alphabet in a span of one minute. As several of the letters are repeated, the scores on this test do not tell us about the number of letters children are familiar with, but the level of automaticity they have established in identifying the letters.

As presented in Table A1, on average, children read about 27 letters aloud in one minute. However, there is tremendous variation, such that children's performance ranged from not identifying a single letter to correctly reading aloud 70 letters in one minute. All children, except one were able to correctly identify one or more letters in the span of 60 seconds. The distribution of scores for letter reading fluency is fairly symmetrical with a slight positive skew (see Fig A1 below)

Table A1: Descriptive statistics for the Swahili reading assessments

	n	Mean	SD	Min	Max
LRF	92	26.59	15.71	0	70
Nonword Reading Efficiency	88	18.53	13.63	0	50
ORF (Passage D)	82	24.87	21.08	0	79.35
ORF (Passage E)	67	24.99	17.72	0	72.35
Comprehension (Passage D)	82	1.85	1.78	0	5
Comprehension (Passage E)	67	3.10	1.94	0	5
Word Reading	108	16.45	5.37	3	24
Sentence Reading	107	0.60	0.16	0.23	1
Maze Test	104	1.99	1.38	0	5

Figure A1: Distribution of scores for Swahili Letter Reading Fluency (n=92)

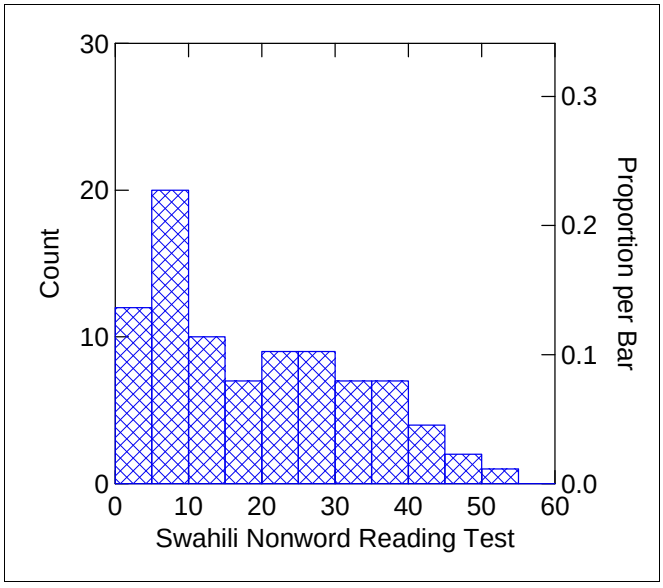


Nonword reading efficiency:

The nonword reading score indicates the efficiency with which children read aloud mono- or bi-syllabic nonwords. The nonwords were constructed in keeping with the phonological rules for Swahili and children's performance on this test indicates phonological recoding ability.

On average, children were able to correctly read about 16 nonwords in the span of one minute. Again there is wide variation with children's performance ranging from no words read correctly to 50 words read correctly. Only five children (5% of the sample) were unable to correctly read any of the nonwords within the span of one minute. The distribution of scores is slightly positively skewed as represented in Fig A2.

Figure A2: Distribution of scores for the Swahili nonword reading efficiency test (n=88)



Oral Reading Fluency (Swahili Passages D and E):

The Swahili oral reading fluency score indicates the speed and accuracy with which children read aloud words in connected text. Scores on this test represent number of words correctly read in the span of one minute. Children were given two passages that were drawn from Grade 1 Swahili primers.

Interestingly, despite the wide variation in oral reading fluency scores, on average for both Passages D and E, children read about 25 words per minute, suggesting that the average fluency scores for both second and third graders as a group is about 25 words per minute. In both cases this is just over half (61%) of the passage.

For Passage D, children's performance ranged from no words read correctly to 79 words read correctly in the span of 60 seconds. For Passage E, children's performance was similar and ranged from no words read correctly to 72 words read correctly in the span of 60 seconds.

The distribution of scores for oral reading fluency has a floor effect for both Swahili passages (see Figs A3 and A4). About 27% (n=22) of the total sample of 82 children were unable to read any of the words of Passage D correctly and 16% (n=11) of the total sample of 67 children were unable to read any of the words of Passage E correctly. These children are henceforth referred to as nonreaders.

Figure A3: Distribution of Oral Reading Fluency scores for Passage D (n=82)

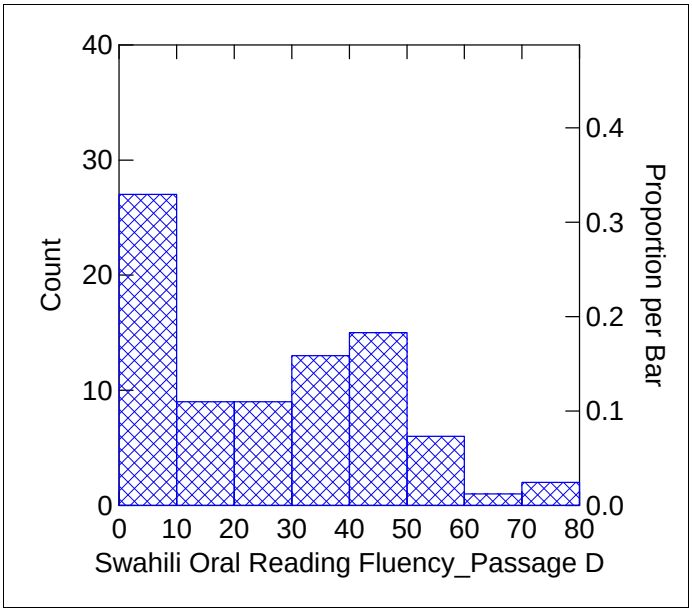
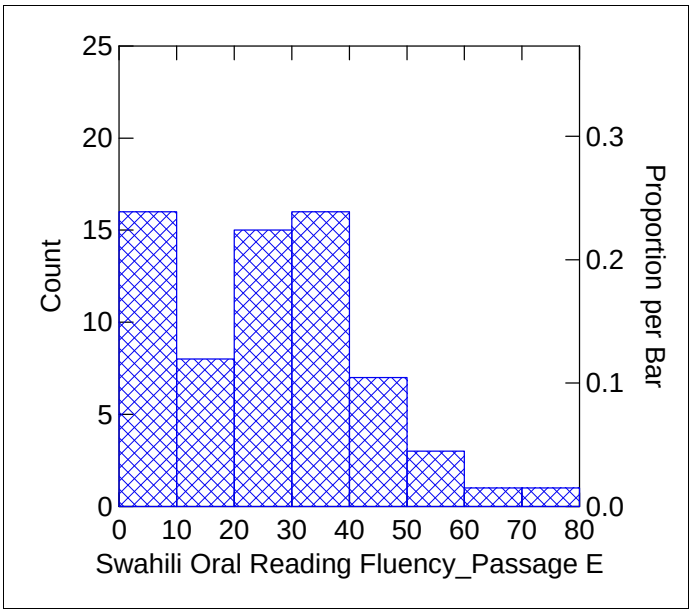


Figure A4: Distribution of ORF scores for Passage E (n=67)



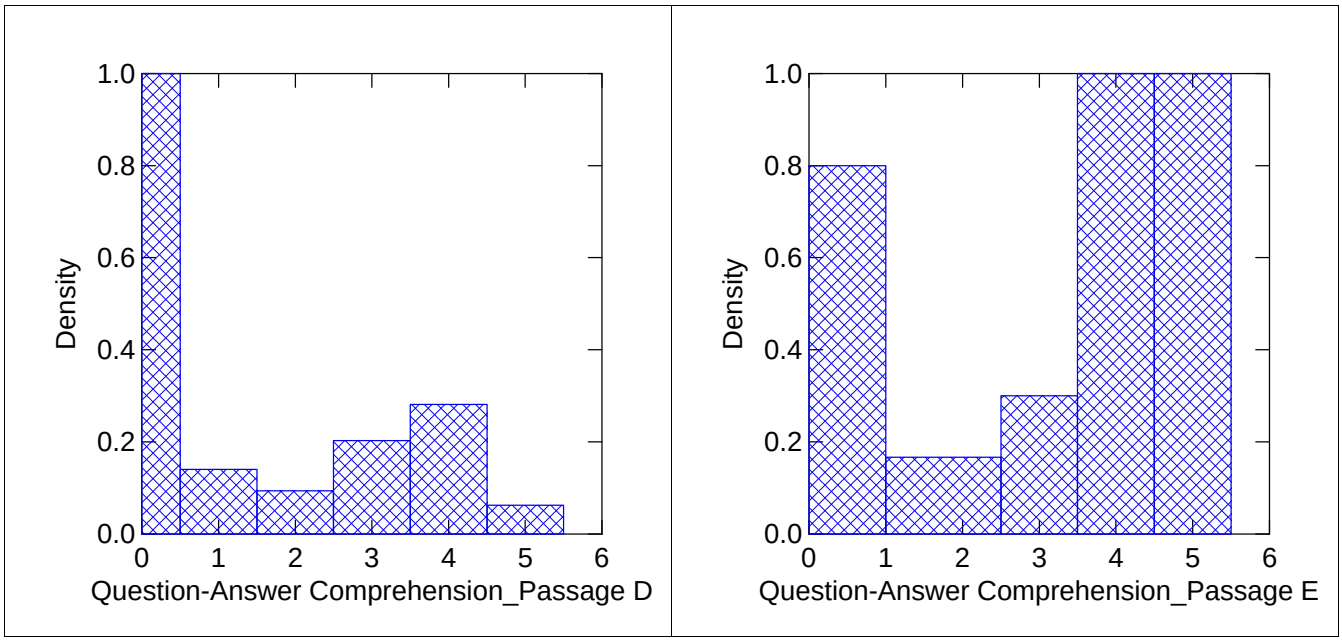
Passage Comprehension (Passages D and E)

Children were orally presented with five questions for each of the two Swahili Passages immediately after they had finished reading the passage.

On average, children were able to correctly respond to 1 question on Passage D and 3 questions on Passage E. Here again, children's performance varied across the full range of the test, i.e. from 0-5 for both passages.

Fig A5 presents the frequency distributions for both Passages. There were fewer children who were unable to respond to any question correctly for Passage E than for Passage D. This is in keeping with the smaller floor effect (16%) for Passage E in comparison to Passage D (26%). Furthermore, 60% of the sample of children were able to respond correctly to 4 or all 5 questions for Passage E, while a fewer number (27%) of children were able to respond correctly to 4 or 5 questions for Passage D. All of this suggests that Passage E was a relatively easier passage in comparison to Passage D, highlighting the difficulty in developing consistent and equivalent tests of reading comprehension. In any case, irrespective of difficulty as noted above children's oral reading fluency rates were similar for both passages - about 25 words read correctly per minute.

Figure A5: Frequency distribution for number of questions responded correctly to questions based on Swahili Passages D (N=82) and E (N=67)



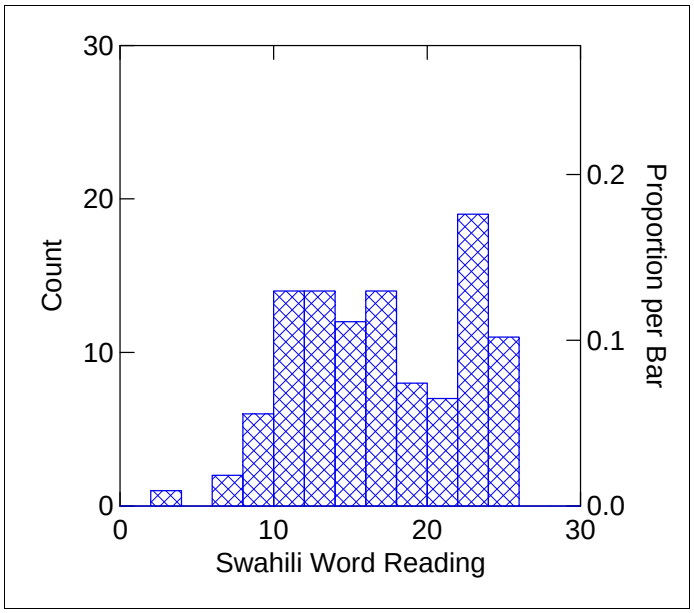
*Note:
Of the 32 children who did not respond correctly to any of the questions for Passage D, 20 children were nonreaders
Of the 16 children who did not respond correctly to any of the questions 11 were nonreaders

Word Reading

This test was group administered and required children to indicate whether the target word was a real word or a nonword. Hence, this test assesses both (a) the ability to decode words, and (b) knowledge of words.

On average, children scored about 17 from a total of 70 words on the word reading test. Here again there was wide variation ranging from 3 to 24. The distribution of scores are slightly negatively skewed with no definite peak (see Fig A6)

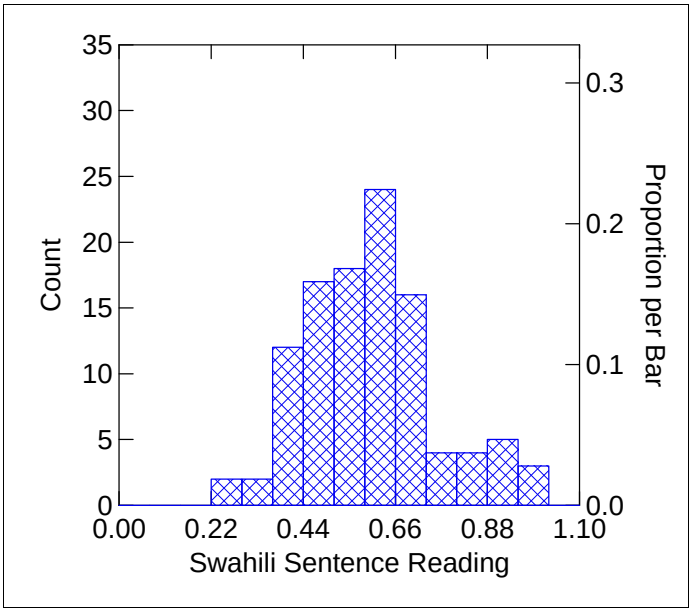
Figure A6: Distribution of scores for the Swahili word reading test (n=108)



Sentence Reading

This test was group administered as well as timed and required children to indicate whether the target sentences were true or false. Hence, this test assesses both (a) the ability to decode words, and (b) sentence comprehension. Since this is a timed test, the scores for this test were calculated to control for guessing and Table A1 presents the descriptive statistics for this score. The distribution of scores for this test is fairly symmetric as represented in Fig A7.

Figure A7: Distribution of scores for the Swahili Sentence Reading test (n=107)

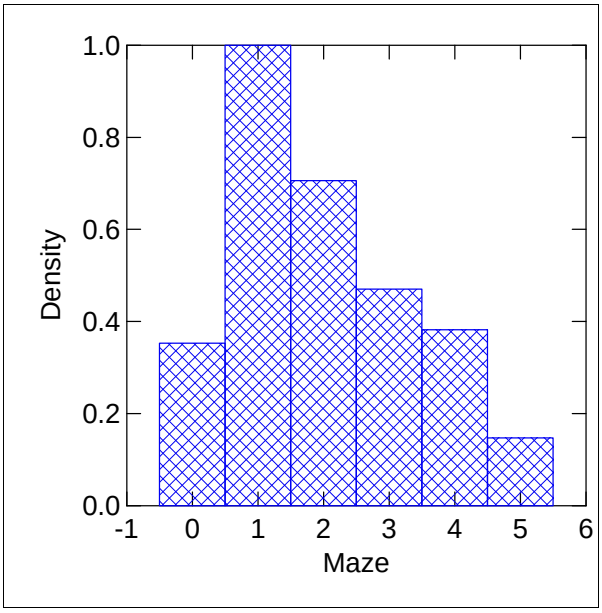


Maze Test

This test requires children to silently read a passage with five target words missing. Children are required to select each of the target words from among three distractor words. Hence, this test assesses (a) children's ability to read words silently in connected text and (b) children's passage comprehension. Scores for this test range from 0 to 5.

As presented in Table A1, on average children were able to correctly respond to 2 items correctly on the Maze task. Children's performance ranged from not getting a single item right to correctly responding to five items. There was a slight floor effect (similar to that of the Passage Reading tests) with about 12% of the sample (n=12) not getting a single item right as represented in Fig A8 below.

Figure A8: Frequency distribution of scores for the Maze test (n=104)



Examining the association between oral reading fluency and comprehension for each Swahili passage:

As presented in Table A2, there is a strong correlation between oral reading fluency and the number of comprehension questions responded to correctly for each of the passages ($r = 0.81$, $p<.0001$ for Passage D and $r = 0.79$, $p<.0001$ for Passage E). In other words, children who read faster and more accurately were more likely to respond correctly to a greater number of the questions based on the passage, suggesting that oral reading fluency is in fact significantly associated with comprehension for Swahili, a regular orthography.

Examining the association of oral reading fluency and comprehension scores across the passages:

There is a very strong correlation between the oral reading fluency scores for the two passages. ($r = 0.93$, $p<.0001$). This suggests that children are likely to maintain their rank ordering in terms of accuracy and speed of reading words across the two passages.

Also, there is a high correlation between children's comprehension scores for the two passages (r=0.82, p<.0001).

Table A2: Examining the association between Swahili Passage D (n=82) and Passage E (n=67)

	ORF (Passage D)	ORF (Passage E)	Comprehension (Passage D)	Comprehension (Passage E)	Grade
ORF (Passage D)	--				
ORF (Passage E)	.93***	--			
Comprehension (Passage D)	.81***	.72***	--		
Comprehension (Passage E)	.76***	.79***	.82***	--	
Grade	.53***	.49***	.53***	.59***	--

Examining the associations between oral reading fluency and comprehension with grade:

The sample comprises 35 second graders and 86 third graders. The correlation estimates as presented in Table A2 indicate that third graders are more likely to read a greater number of words accurately per minute than second graders and that they are also most likely to respond correctly to a greater number of comprehension questions than second graders. These associations are similar for both passages and the correlations range from 0.49 to 0.59. **This suggests that the ORF and question-answer comprehension measures demonstrate developmental sensitivity such that third graders are likely to perform better than second graders.**

Examining letter reading fluency scores for nonreaders:

As a next step the scores for letter reading fluency were examined for children who were identified as nonreaders, i.e. were unable to read any of the words correctly on the Passage Reading assessments. The descriptive statistics presented in Table A3 indicate that there was wide variation among nonreaders in their scores for letter reading fluency. All children were correctly able to identify one or more letters within the span of 60 seconds. For the group of nonreaders based on Passage D, letter reading fluency scores ranged from a low of 2 to a high of 43, and for the group of nonreaders based on Passage E scores ranged from a low of 5 to a high of 43. Although there is wide variation in nonreaders ability levels on this task, none of them reach the maximum fluency levels (score of 70) of children who are classified as readers

These results indicate that letter reading fluency is a necessary but not a sufficient condition for reading words in connected text. However, the letter reading fluency test is useful in providing information on children for whom otherwise we would have no data if only the Passage Reading tests were to be administered.

Table A3: Descriptive statistics for Letter Reading Fluency for nonreaders for Swahili Passages D and E

	n	**Mean**	**SD**	**Min**	**Max**
LRF: Passage D Nonreaders	22	15.64	11.05	2	43
LRF: Passage E Nonreaders	11	17.36	11.94	5	43

Divergent-Convergent Validity:

As presented in Table A4, the ORF scores are positively correlated with all of the reading assessments – Letter Reading Fluency, Nonword Reading Efficiency, Word Reading, and Sentence Reading. These correlations range from a magnitude of 0.67 to 0.75 for Passage D and 0.59 to 0.80 for Passage E.

Interestingly, and as would be expected the ORF scores are most strongly associated with the Passage Comprehension test and the Sentence Reading test than any of the other reading measures. Since the Sentence Reading test, as indicated above assesses children's ability to decode words in connected text as well as comprehension ability at the sentence level. Hence, the stronger correlation of ORF with Sentence Reading and the Question-Answer Comprehension assessments than any of the other assessments further corroborates the use of ORF as an index of children's comprehension ability.

The *Maze task*, on the other hand, has moderate correlations with the ORF and passage comprehension measures at the word, sentence and passage level (see Table A4 and A5). However, like the ORF measure, the Maze test has stronger correlations with assessments of reading connected text and comprehension than with the assessments of letter reading fluency and word reading. Its correlations with the former set of tests ranges from .56 to .69 and for the latter range from .50 to .44. This suggests that the Maze is perhaps a valid measure of children's ability to decode words in connected text and comprehension ability. However, unlike the ORF assessment, the association of the Maze test with the comprehension test and sentence reading test are weaker. This provides further evidence that ORF is a more robust index of children's reading and comprehension ability than Maze.

Table A4: Examining the associations between the reading assessments – Letter Reading Fluency, Word Reading, and Sentence Reading with Oral Reading Fluency based on Passage D (n=68)

	ORF (Passage D)	Comprehension (Passage D)	LRF	Nonword Reading	Word Reading	Sentence Reading
ORF (Passage D)	--					
Comprehension (Passage D)	.79***	--				
LRF	.67***	.63***	--			
Nonword Reading	.79***	.71***	.57***	--		
Word Reading	.67***	.68***	.57***	.63***	--	
Sentence Reading	.76***	.73***	.59***	.77***	.63***	--
Maze	.61***	.69***	.48***	.61***	.44***	.60***

Table A5: Examining the associations between the reading assessments – Letter Reading Fluency, Word Reading, and Sentence Reading with Oral Reading Fluency based on Passage E (n=57)

	ORF (Passage E)	Comprehension (Passage E)	LRF	Nonword Reading	Word Reading	Sentence Reading
ORF (Passage E)	--					
Comprehension (Passage E)	.81***	--				

LRF	.59***	.59***	--			
Nonword Reading	.78***	.69***	.49***	--		
Word Reading	.63***	.71***	.44***	.56***	--	
Sentence Reading	.80***	.78***	.57***	.79***	.68***	--
Maze	.56***	.56***	.47***	.59***	.48***	.64***

Exploring the possibility of creating a composite score based on letter reading fluency, nonword reading efficiency and oral reading fluency: Do these three measures tap the same underlying construct?

Principle component analysis (PCA) was conducted to determine whether letter reading fluency, nonword reading efficiency and oral reading fluency tap the same underlying construct. As there was a very high correlation between the oral reading fluency scores for the two passages ($r = .93$, $p<.0001$) only one of these two passages, Passage D was used in compositing these two measures.

Correlational analysis indicates that the three tasks are fairly strongly correlated with each other and suggests that the three tests may represent a unitary construct.

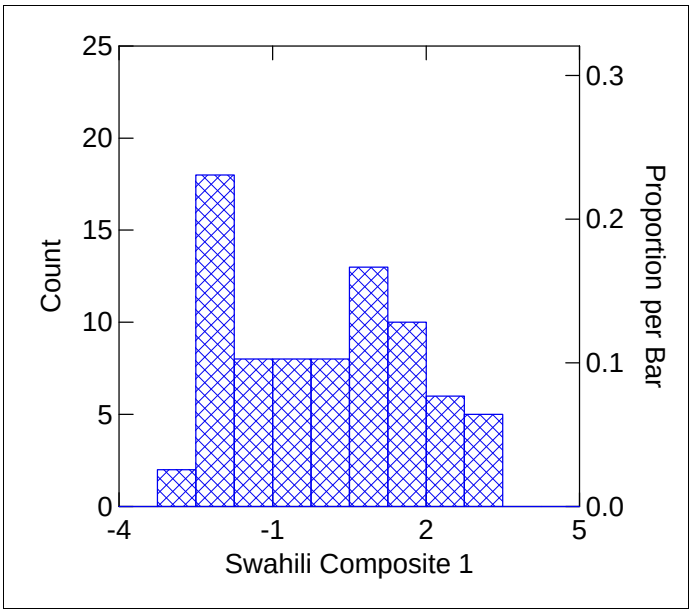
The overall Cronbach's Alpha reliability estimate is quite high at .86 for the three assessments as a group. Furthermore, PCA analysis indicates that the three tests represent a single unitary construct where each test is equally weighted in the creation of the composite. This first composite captures the bulk of the total variation, i.e. 78% of the total variation. The magnitudes associated with each of the tests are .60 for oral reading fluency, .54 for letter reading fluency, and .59 for nonword reading efficiency.

Given that each of the tests carry similar weights, for future work a composite can be created by computing an average of the sum of standardized scores for all three assessments.

Examining the association of the first composite score with the question-answer comprehension test

The univariate distribution of scores for the first composite based on letter reading fluency, nonword reading efficiency, and oral reading fluency as represented in Fig A9 is bimodal, with a first peak at the lower end of the distribution representing children who scored low on the three tests and a second peak around the average value for the composite measure.

Figure A9: Distribution for the composite score based on the three Swahili reading tests (n=78)



The high correlation of the composite measure with the question-answer comprehension test (r = .80, p<.0001) indicates that the composite of the three reading tests is a valid index of children's reading comprehension ability.

A parsimonious representation of early reading ability

Given that both the construction and administration of the nonword reading efficiency test presents several challenges (see discussion in accompanying paper) for cross-linguistic comparisons as well as for large scale assessments, a parsimonious yet efficient representation of early reading ability can be estimated on the basis of the letter reading fluency test and the oral reading fluency test.

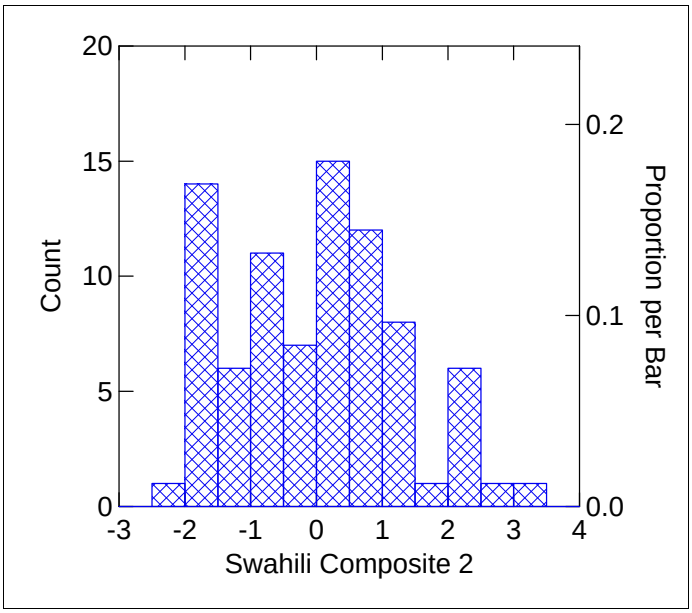
The overall Cronbach's Alpha reliability estimate for the letter reading fluency and the oral reading fluency test is quite high at .79. Here too, PCA analysis indicates that the two tests represent a single unitary construct with each test weighted equally (.71) in the creation of the composite. This composite captures a substantial proportion of the total variation, i.e. 83% of the total variation.

Here too, given that each test is identically weighted in the creation of the composite, for future work a composite can be created by computing an average of the sum of the standardized score for the two assessments.

Examining the association of the second composite score with the question-answer comprehension test

The univariate distribution of scores for the second composite based on letter reading fluency and oral reading fluency as represented in Fig A10 is somewhat symmetrical with wide variation in scores. A high score on this composite represents children who scored high on both the letter reading fluency and the oral reading fluency tests, and vice versa for low scores.

Figure A10: Distribution for the composite score based on the two Swahili reading tests (n=83)



For this second composite too, the moderately high correlation of the composite measure with the question-answer comprehension test ($r = .79$, $p<.0001$) indicates that the composite of the two reading tests alone also functions as a valid index of children's reading comprehension ability.

It is important to note that the use of a composite score will depend on the inference to be drawn. If the focus is on understanding children's general reading ability as defined by these measures than the composite score is a useful and parsimonious representation. On the other hand, if the focus is on understanding children's performance on each of the component measures so as to more effectively inform instructional goals and intervention programs then scores on the individual measures are of greater use. Scores on individual tests are also of greater use to understand children's reading skill levels, especially for children for whom the oral reading fluency measure yields no information.

Summary of Findings for Swahili

- ③ Oral Reading Fluency (ORF) is moderately to strongly correlated with all the reading measures. However, relative to other reading tests, it is most strongly correlated with the Sentence Reading test and the Question-Answer Comprehension test thus corroborating the use of ORF as a valid index of primary grade students' comprehension ability for Swahili, a regular orthography.
- ③ Letter Reading Fluency (LRF) is useful for obtaining information about students at the lower end of the ORF distribution, particularly for students for whom ORF yields no information.
- ③ ORF, Nonword Reading Efficiency and LRF represent a unitary construct of reading ability as suggested by principal components analysis.

③ Strong associations of the composite score with the Question-Answer Comprehension test further corroborates the validity of these tests as an index of reading comprehension ability

Assessing early reading ability in English: Descriptive Statistics

The descriptive statistics for the Swahili reading assessments are presented in Table A6.

English Letter Reading Fluency:

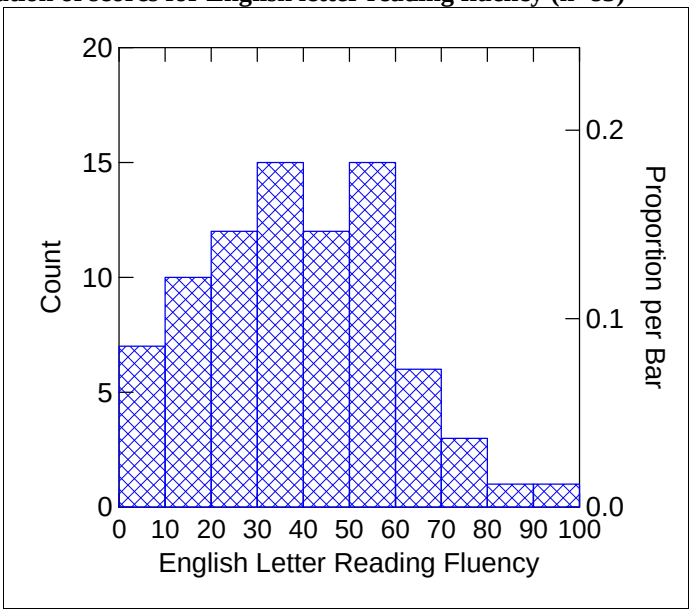
The English letter reading fluency score, like for Swahili indicates the speed and accuracy with which children read aloud randomly arranged lower-case and upper-case letters of the English alphabet in a span of one minute. As several of the letters are repeated, the scores on this test do not tell us about the number of letters children are familiar with, but the level of automaticity they have established in identifying the letters.

As presented in Table A6, on average, children read about 38 letters aloud in one minute. However, there is tremendous variation, such that children's performance ranged from not identifying two letters to correctly reading aloud 90 letters in one minute. In other words, all children were able to correctly identify one or more letters in the span of 60 seconds. The distribution of scores for letter reading fluency is fairly symmetrical with a slight positive skew (see Fig A11 below)

Table A6: Descriptive statistics for the English reading assessments

	n	Mean	SD	Min	Max
LRF	83	37.98	20.00	2	90
Nonword Reading Efficiency	79	17.73	14.43	0	50
ORF (Passage C)	71	24.31	22.85	0	90
ORF (Passage D)	69	24.35	22.05	0	85.5
Comprehension (Passage C)	71	0.69	1.06	0	5
Comprehension (Passage D)	69	0.57	0.79	0	3
Maze Test	110	0.75	0.82	0	3

Figure A11: Distribution of scores for English letter reading fluency (n=83)

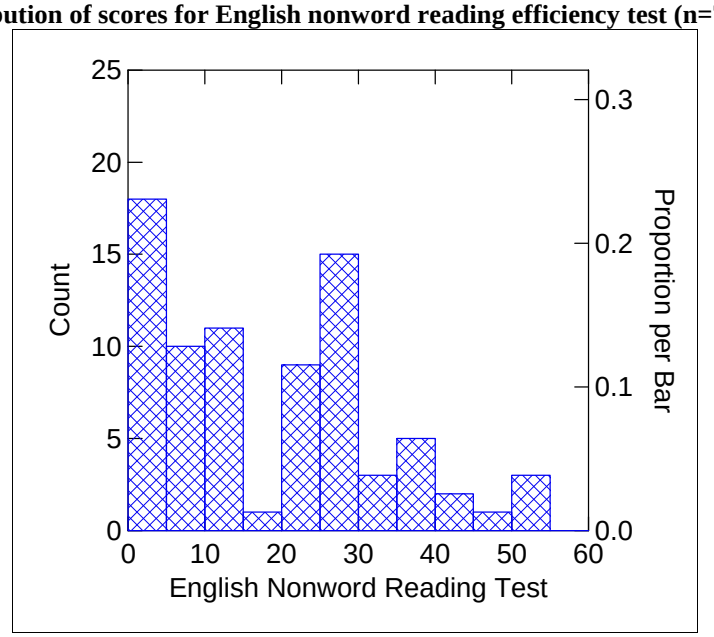


Nonword reading efficiency:

The nonword reading score indicates the efficiency with which children read aloud mono- or bi-syllabic nonwords. The nonwords were constructed so as to not violate the phonological rules for English and children's performance on this test indexes their phonological recoding ability.

On average, children were able to correctly read about 18 nonwords in the span of one minute. Again there is wide variation with children's performance ranging from no words read correctly to 50 words read correctly. Fifteen children, i.e. 18% of the sample were unable to correctly read any of the nonwords within the span of one minute. Given this floor effect, the distribution of scores is spositively skewed as represented in Fig A12.

Figure A12: Distribution of scores for English nonword reading efficiency test (n=79)



Oral Reading Fluency (English Passages C and D):

The English oral reading fluency score indicates the speed and accuracy with which children read aloud words in connected text. Scores on this test represent number of words correctly read in the span of one minute. Children were given two passages that were drawn from their English grade-level primers.

Interestingly, despite the wide variation in oral reading fluency scores, on average, for both Passages C and D, children read about 24 words per minute. Moreover, the ORF rates for English are similar to Swahili for this group of second and third graders. This represents just under half (39% for Passage B and 40% for Passage C) of the passage.

For Passage C, children's performance ranged from no words read correctly to 90 words read correctly in the span of 60 seconds. For Passage E, children's performance was similar and ranged from no words read correctly to 90.5 words read correctly in the span of 60 seconds.

The distribution of scores for oral reading fluency has a floor effect for both English passages (see Figs A13 and A14). About 30% (n=21) of the total sample of 70 children were unable to read any of the words of Passage C correctly and 28% (n=19) of the total sample of 68 children were unable to read any of the words of Passage D correctly. These children are henceforth referred to as nonreaders.

Figure A13: Distribution of Oral Reading Fluency scores for Passage C (n=71)

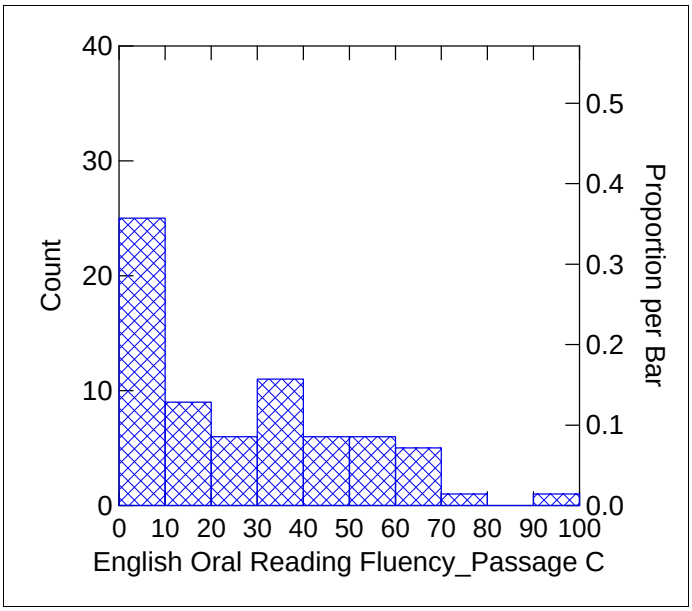
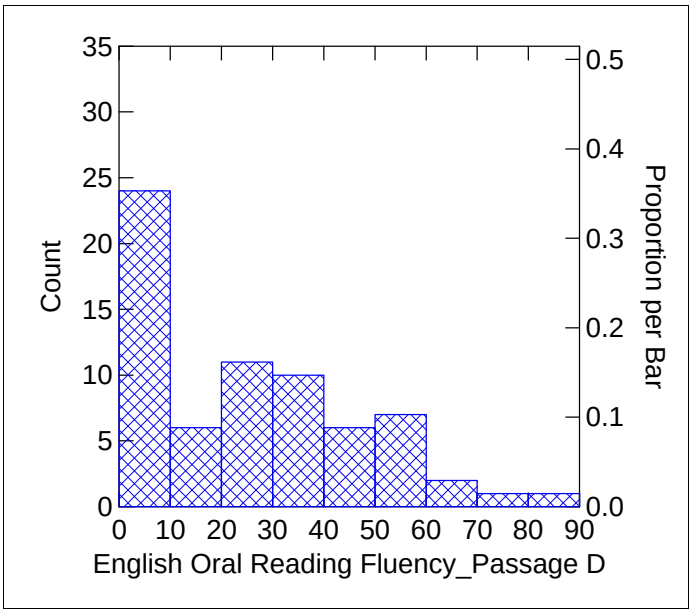


Figure A14: Distribution of ORF scores for Passage E (n=69)



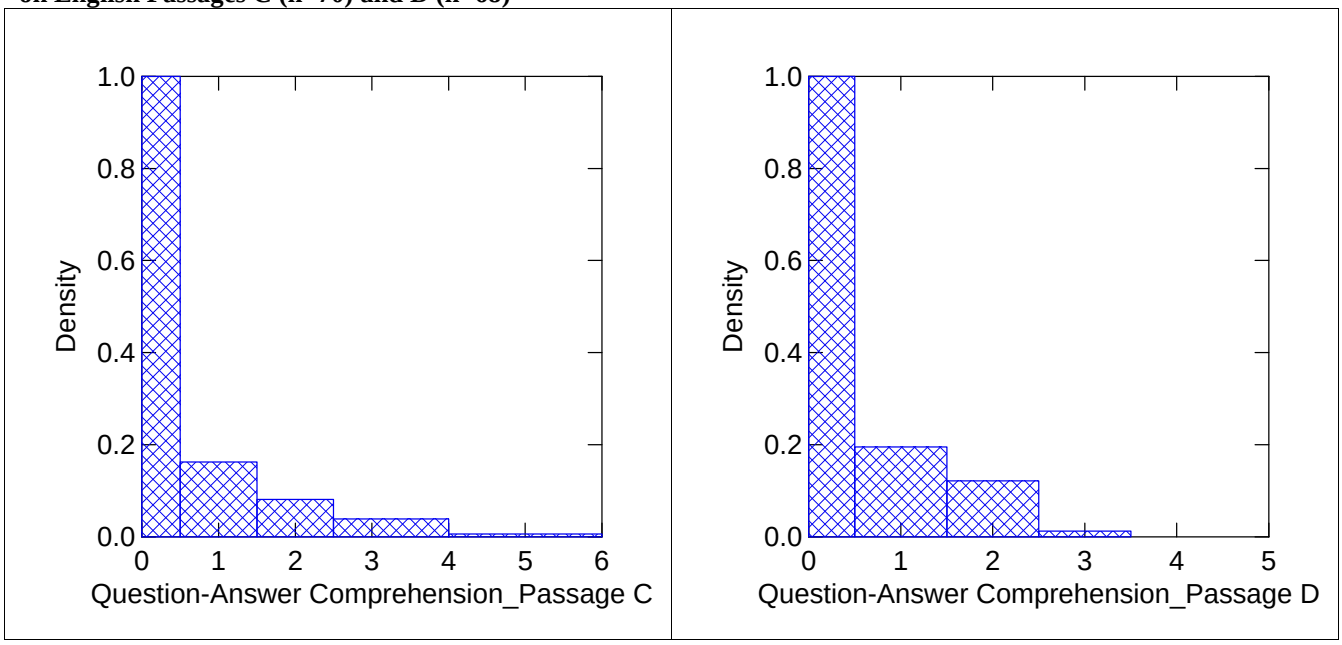
Passage Comprehension (Passages C and D)

Children were orally presented with five questions for each of the two English Passages immediately after they had finished reading the passage.

On average, children were able to correctly respond to 1 question on both Passages C and D. Children's performance varied across the full range of the test, i.e. from 0-5 for Passage C while for Passage none of the children got all the questions right and performance ranged from 0-3.

As represented in Fig A15 there were substantial floor effects for both passages. About 61% (n=43) of the sample of 70 children were unable to respond correctly to any of the questions for Passage C and about 60% (n=41) of the sample of 68 children were unable to respond correctly to any of the questions for Passage D. This is substantially larger than the floor effect for ORF suggesting that not all children who demonstrated some ability to read were able to respond correctly to the comprehension questions, which is further substantiated by the moderate correlations between these two measures for each passage(see below).

Figure A15: Frequency distribution for number of questions responded correctly to questions based on English Passages C (n=70) and D (n=68)

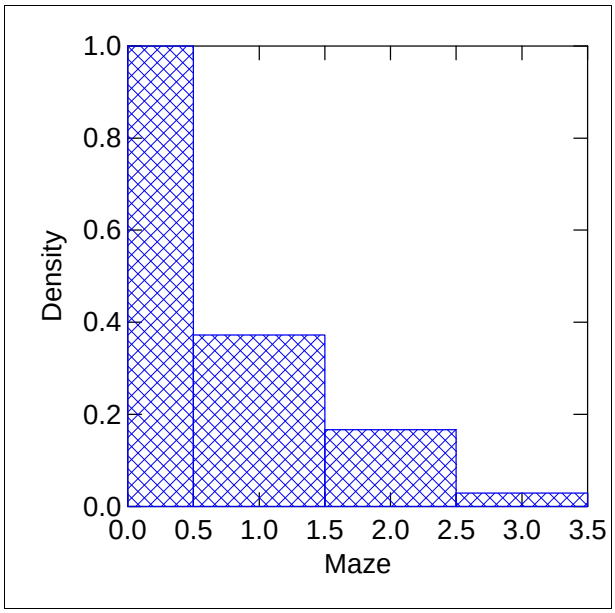


Maze Test

This test required children to silently read a passage with five target words missing. Children were required to select each of the target words from among three distractor words. Hence, this test assesses (a) children's ability to read words silently in connected text and (b) children's passage comprehension. Scores for this test range from 0 to 5.

As presented in Table A6, on average children were able to correctly respond to about 1 item correctly on the Maze task. Children's performance ranged from not getting a single item right to correctly responding to three of the five items. There was a substantial floor effect with about 47% of the children (n=51) from a sample of 109 not getting a single item right as represented in Fig A16 below.

Figure A16: Frequency distribution of scores for the English Maze test (n=109)



Examining the association between oral reading fluency and comprehension for each English passage:

As presented in Table A7, there is a moderately strong correlation between oral reading fluency and the number of comprehension questions responded to correctly for each of the passages ($r = 0.61$, $p<.0001$ for Passage C and $r = 0.55$, $p<.0001$ for Passage D). In other words, children who read faster and more accurately were more likely to respond correctly to a greater number of the questions based on the passage. This association between ORF and comprehension is in keeping with prior research in English. However, the correlations are somewhat attenuated to what has been reported by other researchers for English. This can be perhaps attributed to the second language status of English for this group of Kenyan students.

Examining the association of oral reading fluency and comprehension scores across the passages:

There is a very strong correlation between the oral reading fluency scores for the two passages. ($r = 0.86$, $p<.0001$). This suggests that children are likely to maintain their rank ordering in terms of accuracy and speed of reading words across the two passages. Also, there is a moderate correlation between children's comprehension scores for the two passages ($r=0.51$, $p<.0001$).

Table A7: Examining the association between English Passage C (n=77) and Passage D (n=61)

	ORF (Passage C)	ORF (Passage D)	Comprehension (Passage C)	Comprehension (Passage D)	Maze	Grade
ORF (Passage C)	--					
ORF (Passage D)	.86***	--				
Comprehension (Passage C)	.61***	.70***	--			

Comprehension (Passage D)	.62***	.55***	.51***	--		
Maze	.13	.14	.23~	-.05	--	
Grade	.51***	.45***	.34*	.35**	.19~	--

Examining the associations between oral reading fluency and comprehension with grade:

The sample comprises 35 second graders and 86 third graders. The correlation estimates as presented in Table A7 indicate that third graders are more likely to read a greater number of words accurately per minute than second graders for both Passages C and D. Moreover, on the comprehension test, third graders are more likely to respond correctly to a greater number of comprehension questions than second graders. However, these correlations although similar are low in magnitude. **This suggests that the ORF and question-answer comprehension measures demonstrate some amount of developmental sensitivity such that third graders are likely to perform better than second graders.**

Examining the association of the Maze test with the other reading tests:

The *Maze task*, on the other hand, is not significantly correlated with any of the reading measures (see Table A7). Although the lack of a significant relationship between the English Maze test and the other reading measures may have been affected by (a) the narrow range of the test (0 to 5), (b) the low variability in children's performance on this test (0 to 3) and (c) the substantial floor effects (about 47%), it appears that the English Maze measure unlike for Swahili is not very informative and cannot be recommended as an alternative assessment of children's English reading ability levels.

Examining letter reading fluency scores for nonreaders:

As a next step the scores for letter reading fluency were examined for children who were identified as nonreaders, i.e. were unable to read any of the words correctly on the Passage Reading assessments. The descriptive statistics presented in Table A8 indicate that there was wide variation among nonreaders in their scores for letter reading fluency. All children were correctly able to identify one or more letters within the span of 60 seconds. For the twenty one nonreaders based on Passage D, letter reading fluency scores ranged from a low of 4 to a high of 54, and for the group of nineteen nonreaders based on Passage E scores also ranged from a low of 4 to a high of 54. Although there is wide variation in nonreaders ability levels on this task, none of them reach the maximum fluency levels (score of 90) of children who are classified as readers. Similar to Swahili, these results indicate that letter reading fluency is a necessary but not a sufficient condition for reading words in connected text. However, the letter reading fluency test is useful in providing information on children for whom otherwise we would have no data if only the Passage Reading tests were to be administered.

Table A8: Descriptive statistics for Letter Reading Fluency for nonreaders for English Passages C and D

	n	Mean	SD	Min	Max
LRF: Passage C Nonreaders	21	21.33	13.55	4	54
LRF: Passage	19	22.11	13.67	4	54

D Nonreaders

Exploring the possibility of creating a composite score based on letter reading fluency, nonword reading efficiency and oral reading fluency: Do these three measures tap the same underlying construct?

Principle component analysis (PCA) was conducted to determine whether letter reading fluency, nonword reading efficiency and oral reading fluency tap the same underlying construct for English. As there was a high correlation between the oral reading fluency scores for the two passages ($r = .86$, $p<.0001$) only one of these two passages, Passage C was used in compositing these two measures.

Correlational analysis indicates that the three tasks are fairly strongly correlated with each other and suggests that the three tests may represent a unitary construct (see Table A9).

Table A9: Examining the associations between the English reading tests (n=56)

	LRF	Nonword Reading Efficiency	ORF (Passage C)
LRF	--		
Nonword Reading Efficiency	.87***	--	
ORF (Passage C)	.81***	.88***	--

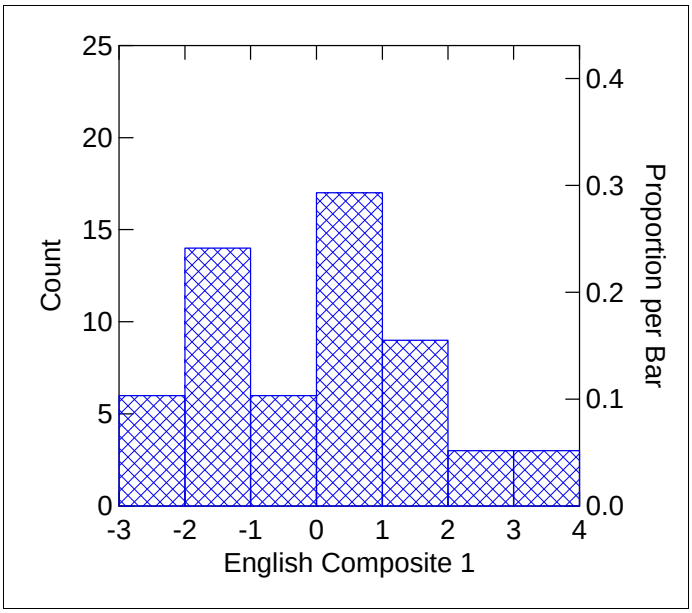
The overall Cronbach's Alpha reliability estimate is quite high at .95 for the three assessments as a group. Furthermore, PCA analysis indicates that the three tests represent a single unitary construct where each test is equally weighted in the creation of the composite. This first composite captures a substantial proportion of the total variation, i.e. 90% of the total variation. The magnitudes associated with each of the tests are .57 for letter reading fluency, .59 for nonword reading efficiency, and .57 for letter reading fluency.

Given that each of the tests carry similar weights, for future work a composite can be created by computing an average of the sum of standardized scores for all three assessments.

Examining the association of the first composite score with the question-answer comprehension test

The univariate distribution of scores for the first composite based on letter reading fluency, nonword reading efficiency, and oral reading fluency as represented in Fig A17 is bimodal, with a first peak at the lower end of the distribution representing children who scored low on the three tests and a second peak around the average value for the composite measure.

Figure A17: Distribution of scores for the first English composite based on Letter Reading Fluency, Nonword Reading Efficiency, and Oral Reading Fluency (n=58)



The moderately high correlation of the composite measure with the question-answer comprehension test ($r = .65$, $p<.0001$) indicates that the composite of the three reading tests is a valid index of children's English reading comprehension ability.

A parsimonious representation of early reading ability

Given that both the construction and administration of the nonword reading efficiency test presents several challenges (see discussion in accompanying paper) for cross-linguistic comparisons as well as for large scale assessments, a parsimonious yet efficient representation of early reading ability can be estimated on the basis of the letter reading fluency test and the oral reading fluency test.

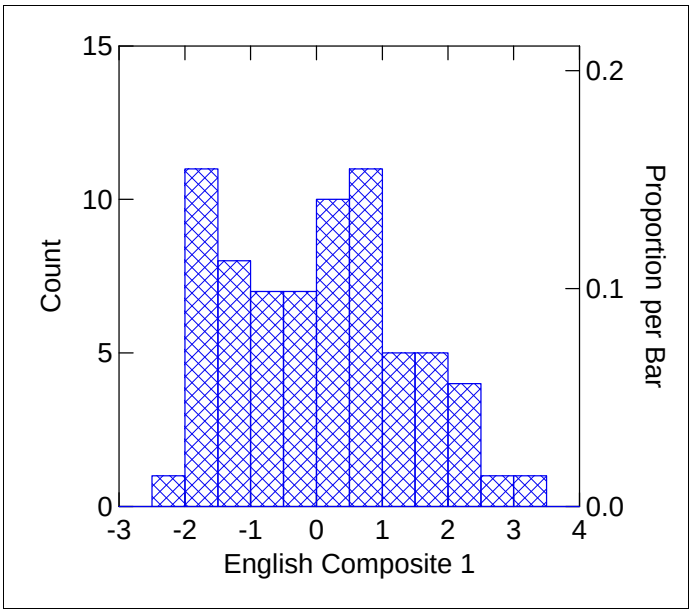
The overall Cronbach's Alpha reliability estimate for the letter reading fluency and the oral reading fluency test is high at .87. Here too, PCA analysis indicates that the two tests represent a single unitary construct with each test weighted equally (.71) in the creation of the composite. This composite captures a substantial proportion of the total variation, i.e. 88% of the total variation.

Here too, given that each test is identically weighted in the creation of the composite, for future work a composite can be created by computing an average of the sum of the standardized score for the two assessments.

Examining the association of the second composite score with the question-answer comprehension test

The univariate distribution of scores for the second composite based on letter reading fluency and oral reading fluency as represented in Fig A18 is also bimodal with wide variation in scores. A high score on this composite represents children who scored high on both the letter reading fluency and the oral reading fluency tests, and vice versa for low scores.

Figure A18: Distribution of scores for the second English composite based on Letter Reading Fluency and Oral Reading Fluency



This second composite has a moderate correlation with the question-answer comprehension test ($r = .60$, $p<.0001$), which is of relatively lower magnitude compared to the first component based on the three reading tests. This suggests that the inclusion of a nonword reading efficiency test will be useful in obtaining a more accurate picture of children's early reading ability. However, the two fluency measures – letter reading fluency and oral reading fluency by themselves also function well as a valid index of primary grade children's reading comprehension ability.

It is important to note that the use of a composite score will depend on the inference to be drawn. If the focus is on understanding children's general reading ability as defined by these measures than the composite score is a useful and parsimonious representation. On the other hand, if the focus is on understanding children's performance on each of the component measures so as to more effectively inform instructional goals and intervention programs then scores on the individual measures are of greater use. Scores on individual tests are also of greater use to understand children's reading skill levels, especially for children for whom the oral reading fluency measure yields no information.

Summary of Findings for English

- ③ Letter Reading Fluency (LRF) is useful for obtaining information about students at the lower end of the ORF distribution, particularly for students for whom ORF yields no information.
- ③ ORF, Nonword Reading Efficiency and LRF represent a unitary construct of reading ability for English as suggested by principal components analysis.
- ③ Moderately strong associations of the composite score with the Question-Answer Comprehension test further corroborates the validity of these tests as an index of reading comprehension ability

Appendix B. Checkpoints for determining the appropriateness of an observational system (Boehm & Weinberg, 1997, p. 111).

1.For what purpose was the system developed?

A. Does the stated purpose match your goal?
B. What will be observed (behavioral definition)?
C. Is the procedure limited by a particular theoretical perspective?

2. Are the conditions for observer reliability met?

A. Behaviors to be viewed are sufficiently specified so as to be:
Mutually exclusive (do not overlap each other). Exhaustive (all behaviors of concern can be classificed, but the need for exhaustive categories depends on the purpose of a particular observation.)
B. Categories are sufficiently narrow so that two or more observers will place an observed behavior into the same category.
C. Is observer interpretation necessary or not?
D. What reliability data are presented?

3. What type of system is employed?

A. Category system: Every unit of behavior observed is categorized into one of the categories specified.
B. Sign system: Selected behavioral units, listed beforehand, may or may not actually be observed during a period of time (such as on a checklist, rating scale, or observation schedule).

4. Are appropriate sampling procedures employed?

A. The procedure for sampling behaviors is systematic: Time sampling-occurrence or nonoccurrence of behaviors within specified uniform time units. Even sampling – event recorded each time it occurs.

B. Is the procedure feasible? How do you sample individuals to be observed? In what period of time? Is the desired detail possible given the number of individuals and time units?
C. What is the coding system like? How complex is the system? Do tallies or codes require memorization? If coding required, is coding indicated on the record form? What is the recording format like?
D. Are the behaviors to be viewed representative? How many behaviors are to be viewed? Over what period of time? Sampling how many children?

5. Are the conditions for validity met?
   A. Are the behaviors you observe relevant to the inferences and interpretations you make?
   B. Have sources of bias been eliminated?
   C. If existing systems are used, are studies available that document the validity of the system?

6. What training procedures are necessary to learn the system?
7. If studies have been reported using the system, what are their outcomes?

8. What modifications will you have to make to the system in which you are interested?

9. Have you tried out the system?
   A. Does it capture the particular behaviors in which you are interested?
   B. How easy is it to use and interpret?
   C. Is it culturally sensitive in regard to the groups with whom you are working?

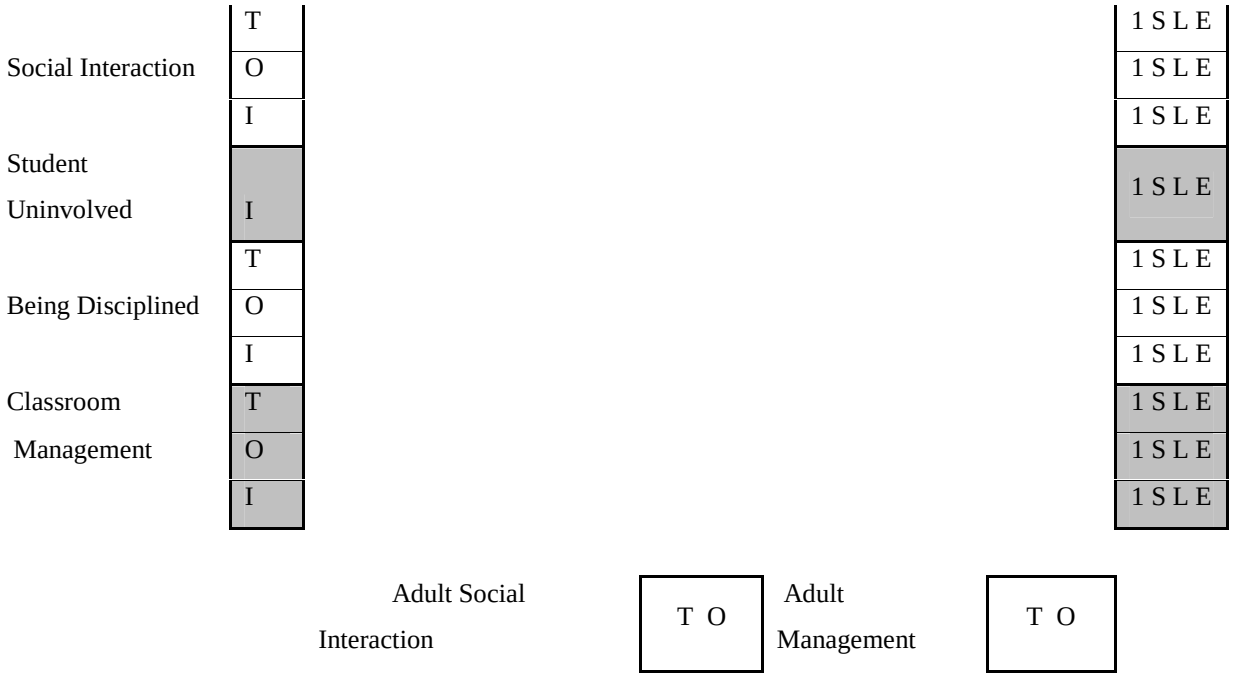
Appendix C. The Stallings Classroom Snapshot

Number of Adults Present ☐

Nunber of Students Present ☐ Males ☐ Females ☐

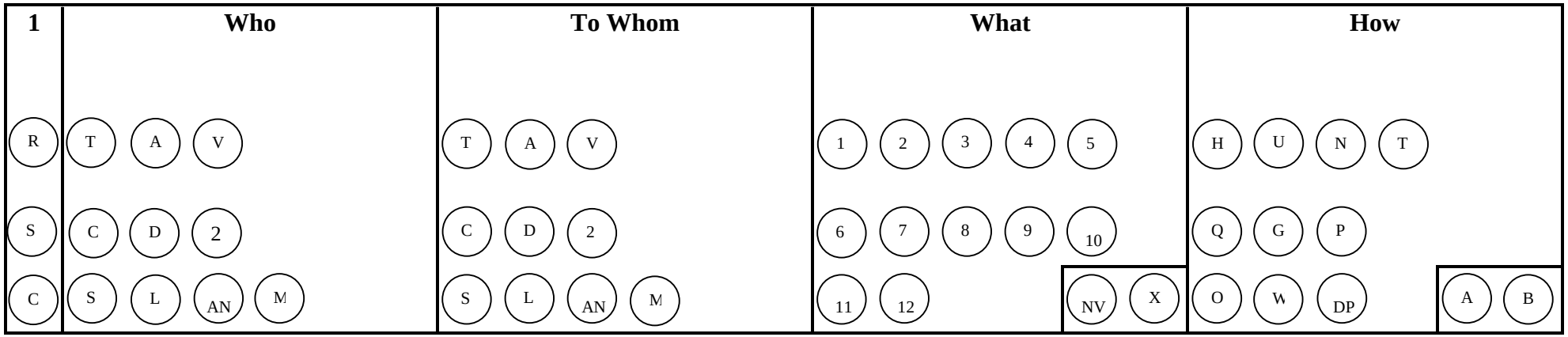
MATERIALS

ACTIVITY		Books	Notebk / Slate	Chalk board	Comput. / Calc.	Manip-ulation	Visual Aids	Co-operative Leaning	None
Receiving Assignments	T	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
	O	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
	I	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
Reading	T	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
	O	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
	I	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
Copying	T	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
	O	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
	I	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
Instruction/ Demonstration	T	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
	O	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
	I	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
Written Assignments/ Seatwork	T	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
	O	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
	I	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
Practice/Drill	T	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
	O	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
	I	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
Discussion	T	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
	O	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
	I	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
Kinesthetic	T	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
	O	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
	I	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
Projects	T	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
	O	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E
	I	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E	1 S L E


T
Social Interaction
O
I
1 S L E
Student
Uninvolved
I
T
Being Disciplined
O
I
Classroom
Management
T
O
I
Adult Social
Interaction
T O
Adult
Management
T O
1 = One Student S = Small Group L = Large Group E =
Everyone

Appendix D. Stallings Five-Minute Interaction: An example of a frame and list of codes **f**rom Stallings (1977) p. 34.

Please note that in the actual instrument, there are multiple frames on one page.



Keys for Codes

Who / To Whom		What		How	
T	Teacher	1	Command or Request	H	Happy
A	Aide	2	Open-ended Question	U	Unhappy
V	Volunteer	3	Response	N	Negative
C	Child	4	Instruction, Explanation	T	Touch
D	Different Child	5	Comments, Greetings; General Action	Q	Question
2	Two Children	6	Task-related Statement	G	Guide / Reason
S	Small Group (3-8)	7	Acknowledge	P	Punish

L	Large Group (9 & up)	8	Praise	O	Object
An	Animal	9	Corrective Feedback	W	Worth
M	Machine	10	No Response	DP	Dramatic Play / Pretend
R	Repeat the frame	11	Waiting	A	Academic
S	Simultaneous action	12	Observing, Listening	B	Behavior
C	Cancel the frame	NV	Nonverbal		
		X	Movement		