Policy Research Working Paper 4812

Measuring Beginner Reading Skills

An Empirical Evaluation of Alternative Instruments and their Potential Use for Policymaking and Accountability in Peru

Ines Kudo
Jorge Bazan

The World Bank
Latin America & the Caribbean Region
Human Development Department
Education Sector
January 2009

Abstract

Based on analysis of reading performance data from 475 third-graders in Peru, this study makes recommendations on improving reading tests, choice of reading standards, and how to present the results at the school and individual levels. The paper reviews the literature on using reading skills measurement in the early grades to guide policymaking, strengthen accountability, and improve education quality. It uses data generated from the same students using two common approaches to measuring reading skills: an individually-administered oral fluency test, and a group-administered written comprehension test designed by the Ministry of Education for the 2006 universal standard test of second grade reading comprehension. These two approaches have sometimes been presented as competing alternatives, but the paper shows that it is better if they are used together, as complements. Based on psychometric analysis, the paper shows that both the oral and written tests adequately measured students' reading abilities. The results show that reading fluency and comprehension are correlated: fluent readers are more likely to understand what they read than non-fluent readers. The strength of the fluency-comprehension relationship depends on the level of fluency, the difficulty of the questions, and social characteristics of the school. The paper recommends using improved versions of both tests to evaluate early grade reading skills, as a central element of a system of accountability for results. It proposes a model for reporting test results desgned to highlight the importance of reading standards, mobilize the education community to reach them, track progress, and identify students in need of extra support.

This paper—a product of the Education Sector, Human Development Department, Latin America & the Caribbean Region—is part of a larger effort in the department to help clarifying standards and establishing a stronger accountability system to improve educational quality and outcomes. Policy Research Working Papers are also posted on the Web at http://econ.worldbank.org. The author may be contacted at ikudo@worldbank.org.

Produced by the Research Support Team

Measuring beginner reading skills:

An empirical evaluation of alternative instruments and their potential use for policymaking and accountability in Peru

Ines Kudo

Jorge Bazan

Acknowledgments

Ines Kudo (corresponding author) is a Senior Education Specialist in the Human Development Department of the World Bank in the Latin America and Caribbean Region. ikudo@worldbank.org. Jorge Bazan is coordinator of the Master's program in statistics of the Pontifical Catholic University of Peru. This study was produced as part of phase three of the World Bank Programmatic Series of Analytical and Advisory Activity, "Recurso Peru", under the supervision and direction of Ian Walker. The authors would like to thank Helen Abadzi, Cristian Aedo, Daniel Cotlear, Luis Crouch, Cecilia Del Solar, and Ian Walker for their insights and valuable contributions to the paper. Our gratitude goes also to Erika Bazan for her support in this process. We would also like to acknowledge the excellent work of the fieldwork coordination team from the Psychology Department at the Pontifical Catholic University of Peru: Cecilia Thorne, Tesania Velázquez and Marcia de la Flor; and the field researchers: Mercedes Ayarza Hernández, Mariella Bazán Maccera, Arturo Cárdenas González, Inés Deustua Rossel, Andrés Figallo Brero, Luis Ernesto Fodale Vargas, Javier Hayashi Hiraoka, Inés María Lazarte Heraud, Diana Loli Valverde, Rosela Millones Cabrera, Eliana Neyra Cruzado, Alberto Piscoya Magallanes, Patricia Ramos Olivera, Jorge Rivas Rivas, Luz Uriol Gárate, María Natividad Vila Torres and Mónica Malpartida.

Abbreviations and acronyms

APAFA	Parents' Association (*Asociación de Padres de Familia*)
CBM	Curriculum-Based Measurment
CNE	National Education Council (*Consejo Nacional de Educación)*
CONEI	Institutional Education Council (*Consejo Educativo Institucional)*
CRM	Criterion-Referenced Measure
CVR	Truth and Reconciliation Commission (*Comisión de la Verdad y Reconciliación*)
DIBELS	Dynamic Indicators of Basic Early Literacy Skills
DRE	Regional Education Office (*Dirección Regional de Educación)*
EBR	Basic Regular Education (*Educación Básica Regular)*
ECE	School Census Evaluation *(Evaluación Censal Escolar)*
EGRA	Early Grade Reading Assessment
ENAHO	National Household Survey (*Encuesta Nacional de Hogares)*
ENCO	National Continuous Survey (*Encuesta Nacional Continua)*
ERDA	Early Reading Diagnostic Assessment
GORT	Gray Oral Reading Test
IE	Education Institution (*Institución Educativa)*
INEI	National Institute of Statistics and Information Technology (*Instituto Nacional de Estadística e Informática)*
INFES	National Institute for Education and Health Infrastructure (*Instituto Nacional de Infraestructura Educativa y de Salud)*
ITBS	Iowa Test of Basic Skills
LLECE	Laboratorio Latinoamericano de Evaluación de la Calidad Educativa (*Latin American Laboratory for Evaluation of Quality of Education)*
MINEDU	Ministry of Education (*Ministerio de Educación)*
MINEDU -UEE	Ministry of Education- Education Statistics Unit *(Ministerio de Educación - Unidad de Estadísticas Educativas)*
MINEDU -UMC	Ministry of Education- Education Quality Measurement Unit (*Ministerio de Educación - Unidad de Medición de la Calidad Educativa)*
MEF	Ministry of Economy and Finances (*Ministerio de Economía y Finanzas)*
MRTA	Tupac Amaru Revolutionary Movement (*Movimiento Revolucionario Túpac Amaru*)
NAEP	National Assessment of National Center for Education
NBI	Basic Unsatisfied Need (*Necesidad Básica Insatisfecha)*
NRM	Norm-Referenced Measure
NRP	National Reading Panel
OECD	Organization for Economic Co-operation and Development
OREALC/UNESCO	UNESCO's Regional Office for Latin America and the Caribbean *(Oficina Regional para America Latina y el Caribe, UNESCO)*
PBI	Gross Domestic Product (*Producto Bruto Interno)*
PCP-SL	Communist Party of Peru - Shining Path (*Partido Comunista del Perú - Sendero Luminoso)*
PEI	Institutional Education Project (*Proyecto Educativo Institucional)*
PISA	Program for International Student Assessment
SAT	Stanford Achievement Test
SBA	Standard-Based Assessment
SDRT	Stanford Diagnostic Reading Test
TPRI	Texas Primary Reading Inventory
UGEL	Local Education Management Unit (*Unidad de Gestión Educativa Local)*
UNESCO	United Nations Education Science and Culture Organization

Contents

Tables

Figures

Measuring beginner reading skills:

An empirical evaluation of alternative instruments and their potential use for policymaking and accountability in Peru

EXECUTIVE SUMMARY

This research was undertaken in the context of a broader body of work (the Peru Recurso 3 Programmatic AAA) aimed at strengthening accountability through the use of results-based standards in education, health and nutrition. In order to improve educational quality and learning outcomes, the Peruvian Government has adopted a strategy focused on clarifying standards and establishing a stronger accountability system, with greater local participation. This study aims to provide a vehicle for The World Bank team to continue the dialogue with the Ministry of Education (MINEDU) regarding how to improve current reading measures, what reading standards are most appropriate, and how the results might be best presented at the school and individual level.

Today, the Government sees universal testing as the best way to provide feedback to policymakers, sector officials, teachers, parents and other stakeholders, in line with its policy of increasing local participation in monitoring service delivery. In 2006, it established a census evaluation system to document the 2nd grade reading performance of each child and school. The ECE (for its initials in Spanish: "Evaluación Censal Escolar") targets the entire second and third grade student population and provides a critical starting point for benchmarking the performance of schools. On the other hand, in several studies and seminars The World Bank has recommended the use of easy-to-administer oral fluency tests to track the students' progress throughout the school year, and as a vehicle to engage parents in monitoring their children's learning. In the past, the two testing approaches have been erroneously taken as alternate, irreconcilable choices, but this paper shows how the two measures can produce helpful, complementary information, finding their combined use more valuable than the choice of either one alone.

Based on a comprehensive literature review and analysis of reading performance data collected in a sample of 475 third-graders in the region of Junin, Peru, this paper stresses the importance of measuring reading skills in the early grades of primary school as a way to produce reliable information that can be used to guide policymaking, strengthen accountability, and ultimately improve quality of education.

The study included an individually administered oral fluency test similar to the one proposed by the Bank in previous studies, and a group-administered written test designed by MINEDU for the 2006 Census Evaluation. Through psychometric analysis, the authors establish that both the oral and written tests adequately measure students' reading abilities, though there is room for improvement of both instruments. For instance, the paper suggests adding more items to the ECE test with increasing degrees of difficulty, making sure that all skill levels are adequately measured and that the test complies with the assumptions of the Rasch measurement model preferred by MINEDU. For the oral test, it recommends using longer passages, asking more questions of varied difficulty, and bilingual testing in the case of non-native Spanish speakers.

Consistent with previous research, our results show that reading fluency and comprehension are positively and significantly correlated with each other. That is, fluent readers are more likely to understand what they read than non-fluent readers. However, results suggest that this relationship is discontinuous and non-linear. Advancing a new analytical approach, the authors find that the strength of this relationship depends on the level of fluency, the difficulty of the questions, and certain social characteristics such as first language of the child.

The study emphasizes the census test's value in helping to establish a baseline over which to set short-term (e.g. annual) attainable goals for improving learning outcomes. In addition, the paper proposes models for reporting future census results at the individual, class, school, district, region and national level. These reports are designed to strengthen horizontal and upward accountability. The proposed models make sure that results are communicated to parents, teachers, principals and the general public in a way that is easy to understand and that facilitates: (a) tracking progress of students and schools relative to previous years; (b) comparing results from schools with similar characteristics or located in the same jurisdictions; (c) comparing results to specific reading goals or standards; and (d) identifying students or schools at risk of falling behind or already performing at a critical level.

The paper also shows that minimum fluency (e.g. 45 words read correctly per minute) is a good predictor of basic, literal comprehension of short passages; while higher levels of fluency are related to the ability to understand longer, more complex readings. Therefore, besides setting annual reading goals for the written test, the paper recommends the use of progress goals throughout the school year, which could be set in terms of fluency and recall comprehension (e.g. 45 words per minute and the child being able to tell what the story was about). At the early stages of learning to read, fluency is crucial and can only be developed by practicing. The test itself, more than a measure, is also an opportunity to practice and to keep the focus on results. It is relatively easy to administer and gives the teachers information on which students require more attention and which ones are making good progress. It also gives parents information on the progress of their children at school. Even an illiterate parent is able to discern whether his or her child can or cannot read fluently, as the sound of a non-fluent reader is unmistakable.

However, it is not advisable, that the test be used to only coach for fluency. First, because it could have an adverse effect in the children, creating too much pressure to read fast and taking away the pleasure of reading. And second, the evidence that coaching in fluency is *causally* related to comprehension is not strong enough, either based on this study or the literature, to encourage fluency coaching as single-minded strategy in improving reading. However, and when used as a marker or proxy, it can help detecting forward movement and to focus attention on results.

In conclusion, it is not only possible but also desirable to use improved versions of both tests as tools to encourage, monitor and evaluate the development of reading skills in the early grades.

The importance of measuring reading skills in the early grades:

How can reading tests help guide policymaking and strengthen accountability?

1. INTRODUCTION

In the past two decades, Peru has focused on improving access to education and has made impressive gains in coverage, meeting the second Millennium Development Goal (MDG) of universal access to primary education and staying well above regional averages in secondary and tertiary coverage. However, learning outcomes are still quite low by international standards and show little tendency to improve (Banco Mundial, 2006).

In order to improve educational quality and outcomes, the Peruvian Government has adopted a strategy focused on clarifying standards and establishing a stronger accountability system with greater local participation. The government sees universal testing as the best way to provide feedback on outcomes to policymakers, sector officials, teachers, parents and other stakeholders, in line with its policy of increasing local participation in monitoring service delivery. In 2006, the Quality Measurement Unit (Unidad de Medición de la Calidad, UMC) established a census assessment system to document the 2nd grade reading performance of each child and school. The ECE (for its initials in Spanish: "Evaluación Censal Escolar") uses a pencil-and-paper reading comprehension test with a multiple-choice format and targets the entire second and third grade student population. It provides a critical starting point for benchmarking the performance of specific schools over the years, as well as against others in their region and nationwide.

In 2007, the World Bank's Human Development Network –at the request of the Regional President of Junin– conducted a study on the provision of education services in this region. As part of this study, over 1000 third-grade students from public primary schools across the region were given the ECE 2006 test and some of them were also tested in oral reading fluency and comprehension (Banco Mundial, 2008). The study was implemented as part of the third phase of the RECURSO series, an analytical program focusing on accountability and the social sectors (education, health and social protection).

Results from this study revealed that academic performance in Junin is similar to national performance: most children read badly, with poor fluency and limited comprehension. Even those students attending urban, full grade schools –which have more resources and better quality education– finish their primary education with a two-year performance delay, as they read at a level expected for fourth grade students. Preliminary findings also indicate that written and oral measures of reading skills give similar results. In other words, children who read more fluently have better chances of understanding what they read. Inequality persists in this context of generalized low performance. The largest performance gaps are found between the Ashaninka indigenous students and the rest, even when controlling for other social factors such as gender, age, location and type of school, and access to preschool education.

Based on a comprehensive literature review and further analysis of the data collected in Junin, this paper aims to contribute to a better understanding of the relationship between fluency and comprehension measures in the early grades of primary school, and their use in policymaking. To this end, the paper addresses the following questions: How accurately do oral and written tests assess reading skills? What is the relationship between fluency and comprehension measures? Could there be more to it than what pairwise correlations have told us so far? And, based on the answers to those questions: How can reading tests be used to guide policymaking and strengthen accountability?

The paper is structured in five sections: introduction, background, fieldwork methodology, findings, and discussion. The background section presents a short review of the relevant literature and current research on what are the main reading skills; the cognitive processes linking reading fluency with comprehension; the teaching approaches that have proven more effective for early reading; the different instruments used to assess reading performance; the correlations between fluency and comprehension measures; what criteria to use when choosing between oral and written reading tests; and the differential behavior of tests and items in different populations. The section finishes with a review of how reading tests are been used in Peru for developing curricular standards and strengthening accountability.

The third section describes the methodology for the fieldwork. It explains the sampling approach used and describes the sample for this paper. Then, it describes the instruments used, the calculation of final reading scores, and the conditions of test administration.

The fourth section reports the results of the analyses. It begins by comparing and evaluating how accurately the oral and written tests measure reading skills. Then, it analyzes the relationship between scores in both tests, exploring signs of discontinuity in the fluency-comprehension relationship at or around the minimum rate of fluency needed for comprehension according to the literature.

The last section discusses the results taking into account the existing body of research, drawing conclusions with regards to the paper's objectives.

2. BACKGROUND

Reading constitutes the foundation of all school learning: a student who cannot read fluently will not be able to fully benefit from what school has to offer and will hardly become autonomous to access the knowledge imparted in it. Thus, experts in Peru and around the world agree that measuring beginning reading skills in the first grades of primary education is key in monitoring and improving the quality of teaching and learning processes (Abadzi, 2006; Banco Mundial, 2006; Crouch, 2006; Crouch, Abadzi, Echegaray, Pasco, & Sampe, 2005; MINEDU-UMC, 2007; Samuels, 2002; Thorne, 2005).

2.1. Developing reading skills

Like in almost any country around the world, Peruvian children begin to learn reading in first grade of primary school. In order to learn to read and write, children must have previously acquired a set of abilities: oral language skills, the symbolic function, the capacity to focus attention and follow instructions, memory, perceptive spatial skills, laterality, and so on (Thorne, 2005). After these abilities are developed, to read efficiently students must apply letter-sound correspondences, blend sounds together to read words, and recognize that some words are irregular. In addition, they must learn that when they do not understand something they read, they can use comprehension and vocabulary strategies to construct meaning from the passage. This is a huge undertaking for both students and teachers (Vaughn & Linan-Thompson, 2004).

a. What are the main early-reading skills?

There are five critical elements of beginning reading: phonemic awareness, phonics, fluency, vocabulary, and text comprehension (Armbruster, Lehr, & Osborn, 2003; Vaughn & Linan-Thompson, 2004). *Phonemic awareness,* is "the ability to notice, think about, and work with the individual sounds in spoken words" (Armbruster, Lehr, & Osborn, 2003, p. 2). The concept of phonemic or phoneme awareness is comprised in the broader term of phonological awareness. Before children learn to read print, they must understand that words are made up of speech sounds: this is phonological awareness. Conversely, phonemic awareness is the understanding that words are made up of phonemes, it entails, for example, the ability to recognize, isolate, and merge individual sounds of language (Wren, 2004). The United States' National Reading Panel (NRP, 2000) found that phonemic awareness improves children's ability to read words and comprehend what they read, and helps them learn to spell. Phonemic awareness is the basis for learning phonics.

Phonics, on the other hand, is the ability to pair sounds (phonemes) with the letters (graphemes) that represent them. Phonics is also a widely used method of teaching children to read and decode words (Abadzi, 2006). Children begin learning to read using phonics usually around the age of five or six. USA-based research has found that phonics instruction benefits all ages in learning to spell in English (NRP, 2000). In the case of some alphabetic languages, such as Spanish, the spelling systems are even simpler because there is nearly a one-to-one correspondence between sounds and the letter patterns that represent them.

Fluency is the ability to read a text accurately and quickly. Fluent readers read aloud effortlessly and with natural expression, as if they are speaking. Readers who have not yet developed fluency read slowly, word by word, in a choppy manner. Reading fluency consists of precision and speed. Precision is the percentage of words read correctly and speed is the number of words read per minute. Thus, a common measure for fluency is the number of words read correctly in one minute. A review of hundreds of empirical studies concluded that fluent readers are able to focus their attention on making connections among the ideas in a text and between these ideas and their background knowledge and, therefore, are able to focus on comprehension. Non-fluent readers, on the other hand, must focus their attention primarily on decoding individual words, and therefore, they have little attention left for comprehending the text (Armbruster, Lehr, & Osborn, 2003; NRP, 2000).

Vocabulary refers in general to "the words we must know to communicate effectively" (Armbruster, Lehr, & Osborn, 2003, p. 34). Specifically, reading vocabulary refers to words we need to know to understand what we read, that is, the words we recognize in print. There is growing scientific consensus that breadth of vocabulary increases comprehension and facilitates further learning (Hirsch Jr., 2003). Vocabulary experts agree that adequate reading comprehension depends on a person already knowing between 90 and 95 percent of the words in a text (Nagy & Scott, 2000). Knowing that percentage of words allows the reader to get the main thrust of what is being said and therefore to guess correctly what the unfamiliar words probably mean (Hirsch Jr., 2003).

Ultimately, the reason for reading is *comprehension*, which is defined as the process through which meaning is extracted from the written language. There are three different types of comprehension. *Literal* comprehension centers on the recognition or evocation of primary details, main ideas, sequences, or cause-effect, from information that is explicit in the text. *Inferential* comprehension requires making logical connections among facts in texts, deducing events, making generalizations, interpreting facts, and relating the content to previous knowledge or personal experiences, from information implicit in the text (Camba, 2006). Last, *critical* comprehension requires readers to use analysis or make value judgments about what was read.

Phonemic awareness and phonics contribute in the process of learning to decode texts, while fluency and vocabulary contribute to understanding what is being decoded. But if children cannot understand what they read, then they are not really reading (Armbruster, Lehr, & Osborn, 2003).

b. Why are reading comprehension and fluency related?

LaBerge and Samuels' theory of automaticity argues that reading words fluently must occur before comprehension is possible. The underlying premise is as follows: "It is assumed that we can only attend to one thing at a time, but we may be able to process many things at a time so long as no more than one requires attention" (LaBerge & Samuels, 1974, p. 295). The authors proposed that reading involves the coordination of several simultaneous processes. When a skill is practiced enough, it becomes automatic, and when this happens, it requires little to no attention. Hence decoding letters must be automatic to read words, and reading words must be automatic in order for attention to be used for comprehension.

Similarly, more recent theories suggest that when reading a passage, if decoding does not happen quickly, the decoded material will be forgotten before it is understood (Hirsch Jr., 2003). In order to understand a given text the reader needs to be able to decode it fast enough to keep the entire sequence within the confines of the working memory so that the brain can extract meaning (Abadzi, 2006).

Working memory –formerly referred to as short-term memory– is a limited-capacity store for holding temporary information (Jones, 2002). To be registered in permanent memory, the information must first go through the working memory. The latter holds items that come from the outside world or are retrieved from the long-term memory, and sorts them into categories so that the brain can make sense of them. However, the brain allots very little time to working memory, after which, if meaning is not extracted, the material flees the mind, leaving room for the next "batch" (Abadzi, 2006).[1] Experiments on beginning reading have shown that once children master the decoding process, then their working memory frees space to concentrate on comprehension of meaning (Hirsch Jr., 2003; Hudson, Lane, & Pullen, 2005). This is why fluency is so important for comprehension.

---

[1] The first quantification of the short-term memory capacity limit was the "magical" number seven, introduced by cognitive psychologist George Miller (1956). Miller noticed that the memory span of young adults was around seven elements, plus or minus two, regardless of whether they were digits, letters, words, or other units. Later research revealed that there are several factors affecting a person's measured memory span (see Abadzi, 2006; Cowan, 2001; Hulme, Roodenrys, Brown, & Mercer, 1995; Jones, 2002). In general, memory span for verbal contents strongly depends on the time it takes to speak the contents out loud, and on whether the contents are words known to the person or not. Memory span has been found to be around seven for digits, six for letters, and five for words, and it is lower for long words than for short words (Hulme, Roodenrys, Brown, & Mercer, 1995). According to Cowan (2001), working memory has a capacity of about four chunks in young adults and less in children and older adults. A chunk is a small set of items having a common, strong association with each other, and a much weaker one to items in other chunks (Jones, 2002). In a sense, reading is a process by which letters are chunked into words and words are chuncked into structures that are long enough to extract meaning from them, thereby optimizing the use of working memory.

Working memory capacity increases as children grow and their brains develop. Between ages 4 and 5, the neural circuits used in reading are already functional, and by age 6 children have the working memory components found in adults. Each component continues its expansion until reaching ages 16 to 19. In the case of malnourished children, this neurological and cognitive development would likely be delayed or impaired. On the other hand, schooling lengthens working memory, and longer schooling is related to a larger capacity (Abadzi, 2006). Prior knowledge speeds up basic comprehension and frees more working memory to make connections between the new material and previously learned information, to draw inferences, and to ponder implications (Hirsch Jr., 2003). Students can also overcome the limitations of working memory by grasping what kind of text this is, rapidly identifying words, and understanding the grammatical connections between them at the basic level of the sentence. This kind of fluency at the sentence level increases with practice, vocabulary knowledge and exposure to different kinds of writing (Hirsch Jr., 2003).

c. What strategies work best for teaching reading?

In the first years of schooling, children will gradually acquire the skills they need to read proficiently. Teachers can introduce these skills separately, but they should be integrated as quickly as possible. Each of the five main early-reading skills should be emphasized at different times, according to the grade and the psychological development of the child, but they should be included on a daily basis as appropriate for the grade (Vaughn & Linan-Thompson, 2004).

For instance, the focus in kindergarten in many countries is on *phonemic awareness*. Instruction in phonemic awareness involves teaching children to "focus on and manipulate phonemes in spoken syllables and words" (NRP, 2000) and should be taught for approximately 15 to 20 minutes daily for instruction to be effective (Vaughn & Linan-Thompson, 2004). Phonemic awareness involves teaching children to blend or segment the sounds in words using letters and to manipulate sounds in speech without any letters as well (NRP, 2000).

In first grade the focus often turns to *phonics* and *word study*, which provide the foundations for literacy instruction in this grade, but related skills such as print awareness and alphabetic knowledge and understanding, can be emphasized since kindergarten. Students should learn to read decodable text in first grade, since doing so allows them to apply their other skills to reading as well. Because of the complexity of English spelling, there has been controversy around the use of phonics in teaching beginning reading since the nineteenth century, but the most recent research, examined by the National Reading Panel (NRP, 2000) indicates that teaching phonics (and related phonics skills, such as phonemic awareness) is a more effective way to teach children early reading skills than is embedded phonics or no phonics instruction. Phonics instruction requires the teacher to provide students with a core body of information about phonics rules, or patterns.[2]

According to research, *fluency* instruction should start the second semester of first grade, when students have built a strong foundation in word identification (Vaughn & Linan-Thompson, 2004). For fluency instruction in first and second grade, reading must be taught in a known

[2] The National Reading Panel (NRP, 2000) indentifies five Phonics Instructional Approaches: (1) Analogy Phonics, or teaching students unfamiliar words by analogy to known words; (2) Analytic Phonics, or teaching students to analyze letter-sound relations in previously learned words to avoid pronouncing sounds in isolation; (3) Embedded Phonics, or teaching students phonics skills by embedding phonics instruction in text reading; (4) Phonics through Spelling, or teaching students to segment words into phonemes and to select letters for those phonemes; and Synthetic Phonics, or teaching students explicitly to convert letters into sounds (phonemes) and then blend the sounds to form recognizable words.

language to take advantage of the word superiority effect. Teachers must devote most of the class hours in grades 1–2 to reading, giving lots of practice and feedback to students, and teaching vocabulary so that children can comprehend as speed increases. Interesting stories in simple language are likely to help struggling students persevere (Abadzi, 2006). Other research has found that in grades 2 and 3 greater fluency growth is positively associated with small-group instruction, and negatively associated with teaching phonics. Taylor et al (2002) found that children in first grade grew more in comprehension and fluency when their teachers asked more high-level questions than other teachers. Analysis was done using Hierarchical Linear Modeling (HLM) to focus on classroom variables. Teachers who were more often observed teaching their students in small groups in first grade also had students who showed larger gains in fluency during the year.

Two instructional approaches have typically been used to teach reading fluency. Guided, repeated oral reading encourages students to read passages orally with systematic and explicit guidance and feedback from the teacher. The other approach is independent silent reading, which encourages students to read silently on their own, inside and outside the classroom, with minimal guidance or feedback. The NPR research (2000) concluded that “guided repeated oral reading procedures that included guidance from teachers, peers, or parents had a significant and positive impact on word recognition, fluency, and comprehension across a range of grade levels” (p.12). In contrast, they were unable to find a positive relationship between independent-reading programs and improvements in reading achievement, including fluency.

Vocabulary should be an ongoing part of the instructional day in kindergarten through third grade, and taught in the context of reading activities (Abadzi, 2006, p. 42). It has long been known that the growth of word knowledge is slow and incremental, requiring multiple exposures to words. One doesn’t just learn a word’s meaning and then have the word. One gradually learns the word’s denotations and connotations and its modes of use little by little over many, many language experiences (Hirsch Jr., 2003; Nagy & Scott, 2000). Armbruster, Lehr, and Osborne (2003) concluded that children learn the meanings of most words indirectly, through everyday experiences with oral and written language: by engaging daily in oral language, by listening to adults read to them, or by reading extensively on their own. Children who were exposed to greater communicational experiences –e.g. those who have better educated parents, are often read aloud to, attend preschool programs– naturally learned more words and have an advantage over those who did not have that chance. As a result, children arrive to school with vastly different vocabularies: children in middle and upper class urban settings enter primary school with a useable vocabulary that is several times larger than that of rural children. For example, research in Ecuador showed that six year-olds in the poorest quartile had and average score of 63 in the *Test de Vocabulario en Imagenes Peabody* (TVIP), while their peers in the wealthier quartile had scored above 100 (Paxson & Schady, 2005).

In vocabulary acquisition, a small early advantage grows into a much bigger one with time: a high-performing first-grader knows about twice as many words as a low-performing one and, as these students go through the next grades, the differential gets magnified (Graves, Brunetti, & Slater, 1982). By 12th grade, the high performer knows about four times as many words as the low performer (Smith, 1941). Thus, if explicit vocabulary development is not included, rural, poor children will continue to be disadvantaged relative to urban, middle class children.

Students also learn vocabulary through direct instruction, such as key word, repeated multiple readings, rich contexts, computer based and pre-instruction, among others. Research suggest that explicit vocabulary instruction methods improve vocabulary knowledge and reading comprehension and the effects are greatest for students with low initial vocabulary knowledge levels (Nelson & Stage, 2007).

Text comprehension is a complex cognitive process. The NRP research (NRP, 2000) concluded that vocabulary development and vocabulary instruction play a very important role in the understanding of what has been read; it is also an active process that requires an intentional and thoughtful interaction between the reader and the text; and finally, the preparation of teachers to better equip students to develop and apply reading comprehension strategies to enhance understanding is intimately linked to students' achievement in this area. They found that teaching a combination of reading comprehension techniques (comprehension monitoring, cooperative learning, use of graphic and semantic organizers, question answering, question generation, story structure, and summarization) is the most effective method to enhance text comprehension, and that when used in combination, these techniques can improve results in standardized comprehension tests.

Taylor's (2002) results for grades 2 and 3 revealed that students had higher growth in reading comprehension when teachers used less teacher-directed approaches (telling or recitation). In grades 4 through 6, asking students higher-level questions after reading is related to their growth in fluency, comprehension, and writing. Coaching in word recognition strategies during reading was related to students' growth in fluency and writing, but not in comprehension. Having students engaged in active responding was also related to their growth in fluency. Teaching with a highly teacher-directed stance and providing whole-class or large-group instruction were negatively related to growth in fluency, comprehension, and writing (B. M. Taylor, Peterson, Rodriguez, & Pearson, 2002).

2.2. Assessing reading skills

There are dozens of tests currently in the market that assess reading skills (see Annex 7.1 for a detailed list). Thus, this is not meant to be an extensive review of all these measures, but only focus on the general approaches to assessing key early-reading skills, discussing examples of oral fluency and written comprehension measures, the relation between the two, and the criteria to choose the most appropriate one.

a. Common measures of early reading skills

As noted before, there are five key beginning reading skills: phonemic awareness, phonics, fluency, vocabulary, and text comprehension. *Phonemic awareness* is usually assessed by examining a student's knowledge of how sounds make words. Common measures of phonemic awareness ask the student to: (1) break spoken words into parts, counting the number of phonemes; (2) blend spoken parts of a word into one word; or (3) delete or add a phoneme to make a new word (Wren, 2004).

Phonics is assessed only with oral instruments, using for example Letter Recognition, Pseudoword Decoding and Word Reading. In the Letter Recognition subtests (Grades K, 1), the student must identify letters of the alphabet by name. In the Pseudoword Decoding or Nonsense Words subtests (Grades 1, 2), the student must correctly identify nonsense words from a list, which assesses knowledge of letter-sound correspondences as well the ability to blend letters together to form unfamiliar "nonsense." The Word Reading subtests (Grades 1, 2, 3) require the student to correctly identify words in a sight-word list (The Access Center, 2005; Wren, 2004).

Phonemic assessment and phonics are generally evaluated through one-on-one oral tests.[3] When timed, such tests also detect naming fluency.

Vocabulary is often measured indirectly through general comprehension subtests, but there are also direct measures. For instance, asking the student to provide a name for pictures (expressive vocabulary), or to match spoken words with pictures (receptive vocabulary); to provide a word that best matches a definition presented verbally by the teacher (expressive vocabulary); to provide a definition to a word (receptive vocabulary); to select a word that does not belong in a group of words; to provide a synonym or an antonym for words; and so on. All these can be evaluated either orally or with written tests.[4]

The most frequent way of measuring *fluency* is to ask a student to read a passage aloud for one minute. The number of correct words read is counted and this total equals a student's oral reading fluency rate (Hudson, Lane, & Pullen, 2005; RTI International, 2008).[5] Abadzi (2006) sets the cutting point between minimally fluent readers and non-fluent readers at around 45 words per minute.[6] At that speed, reading is still choppy but allows understanding, falling probably on Level 3 in the NAEP Oral Reading Fluency Scale (see Table 1), which is already considered fluent.

The NAEP Fluency Scale, in addition to calculating the correct words read per minute, also provides a descriptive guide for oral reading performance based on the student's "naturalness" of reading. The student's performance is rated on a four-point scale, with emphasis placed on phrasing of words, adherence to syntax, and expressiveness (Hudson, Lane, & Pullen, 2005). The National Reading Panel sets fluency at a reading rate faster than 90 words per minute and with expression (Armbruster, Lehr, & Osborn, 2003), closer to level 4 in the NAEP scale.

---

3 Some examples of one-on-one oral assessment of phonemic awareness are: the Comprehensive Test of Phonological Processing (CTOPP), the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), the Developmental Reading Assessment, K-3 - 2nd Edition (DRA-2, K-3), the Early Reading Diagnostic Assessment (ERDA), the Phonological Awareness Test (PAT), and the Texas Primary Reading Inventory (TPRI). There are not many group-administrated, written assessments of phonological awareness, one rare example being the Iowa Test of Basic Skills (ITBS). Examples of oral assessment that measure phonics are DIBELS, ERDA, the Early-Grade Reading Assessment (EGRA), the Test of Word Reading Efficiency (TOWRE), and TPRI (Torgesen, Wagner, & Rashotte, 1999).

4 Examples of tests that measure vocabulary are: the Abecedarian Reading Assessment, Aprenda: La Prueba de Logros en Espanol - 3rd Edition, Assessment of Literacy and Language (ALL); Boehm Test of Basic Concepts - 3rd Edition (Boehm-3); Bracken Basic Concept Scale – Revised; Bracken School Readiness Assessment (BSRA); Brigance Diagnostic Comprehensive Inventory of Basic Skills - Revised (CIBS-R); Brigance Diagnostic Inventory of Early Development II (IED-II), Brigance Screen tests; Developmental Reading Assessment, K-3 - 2nd Edition (DRA-2, K-3) A Spanish version of the DRA-2 is also available - the Evaluacion del Desarrollo de la Lectura (EDL); Early Reading Diagnostic Assessment (EDRA); Gray Diagnostic Reading Tests; Stanford Diagnostic Reading Test -- 4th Edition (SDRT-4); Standardized Reading Inventory - 2nd Edition (SRI-2); Stanford Achievement Test -- 10th Edition (SAT-10); Stanford Reading First; Test of Reading Comprehension -- 3rd Edition (TORC-3).

5 Examples of measures of oral fluency skills are: AIMSweb Standard Reading, DIBELS, EDRA, EGRA, Gray Oral Reading Test IV (GORT - 4), the National Assessment of National Center for Education (NAEP) Fluency Scale, and the Reading Fluency Monitor Read Naturally

6 This is meant to set a line between children who are able to read with enough precision and speed to understand what they read, and those that still stumble with print, not a standard.

Table 1.NAEP Oral Reading Fluency Scale, Grade 4: 2002

Fluent	**Level 4**	Reads primarily in larger, meaningful phrase groups. Although some regressions, repetitions, and deviations from text may be present, these do not appear to detract from the overall structure of the story. Preservation of the author's syntax is consistent. Some or most of the story is read with expressive interpretation.
	Level 3	Reads primarily in three- or four-word phrase groups. Some small groupings may be present. However, the majority of phrasing seems appropriate and preserves the syntax of the author. Little or no expressive interpretation is present.
Non fluent	**Level 2**	Reads primarily in two-word phrases with some three- or four-word groupings. Some word-by-word reading may be present. Word groupings may seem awkward and unrelated to larger context of sentence or passage.
	Level 1	Reads primarily word-by-word. Occasional two-word or three-word phrases may occur—but these are infrequent and/or they do not preserve meaningful syntax.

SOURCE: U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2002 Oral Reading Study

Finally, *reading comprehension* is the most common measure of reading performance, and yet, comprehension assessment is somewhat a controversial topic, in that no general agreement exists on how to do it best (Hess, 2006). *Question answering* is a widely used reading comprehension assessment in classrooms, and it is incorporated frequently within standardized tests. For each passage read, students provide answers to short-answer questions, which could be literal (factual questions about the text), inferential (making logical connections among facts in text or drawing from ideas implicit in the text), or critical (using analysis or making value judgments about what was read). Comprehension can also be tested with *passage recall*, a well-established method that requires students to retell the story in their own words after reading it. Some authors suggest putting a time limit to this process, such as four or five minutes (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Wren, 2004). A third type of comprehension measure involves the student filling in missing words from a passage. This technique, created by Taylor (1953), is widely known as *cloze* or maze. Finally, a recent innovation asks students to read sentences and state whether they make sense or not. A sensible sentence might read, "The man eats the apple." A nonsense sentence might read, "The chair was hungry."

b. Silent "pencil-and-paper" measures of reading comprehension

Most national and international assessments use pencil-and-paper tests, and almost all comprehensive tests measure reading comprehension. In the United States, some tests that include reading comprehension measures are the Comprehensive Test of Basic Skills (CTBS), the Colorado Student Assessment Program (CSAP), the Iowa Test of Basic Skills (ITBS), the Washington Assessment of Student Learning (WASL), and the Stanford Achievement Test -- 10th Edition (SAT-10).

In Peru, the Quality Measurement Unit (Unidad de Medición de la Calidad, UMC) at the MINEDU has carried out a series of survey-based national assessments in various subject areas and grades, including reading. In 2006, it established a universal assessment system to document the 2nd grade reading performance of each child and school. Although the first ECE (for its initials in Spanish: “Evaluación Censal Escolar”) assessment was organized on a very tight time-line, it reached around 55 percent of the target. The ECE-2007 showed great progress in terms of coverage, reaching over 90 percent of the target, and included a test on math reasoning. The ECE provides a critical starting point for benchmarking the performance of specific schools against others in their region and nationwide.

ECE tests are formulated based on the National Curricular Design (DCN), taking into account the learning achievements expected for the end of second grade in the area of integral communication. Because all DCN’s goals for second grade on integral communication refer to reading comprehension, the test has focused exclusively on this skill.

The test follows a multiple-choice format, with three alternative answers to choose from, and questions have different degrees of difficulty: word comprehension, sentence comprehension, and connected-text comprehension with literal and inferential questions. Each question provides information on the development of the pertinent abilities and allows the identification of the reading tasks that the students master. The MINEDU defines “reading tasks” as the cognitive processes students use to understand a written text. This includes identifying the main event in written news, recognizing the sequence of facts in a story, and so on.

The ECE classifies second grade students into four groups based on three specific attainment levels (0, 1, 2 and 3) with clear, meanings. A student in the Group 0 cannot read and comprehend even the simplest sentence. Reaching attainment Level 1 supposes the ability to read and comprehend a simple sentence. Level 2 means that the child can read a more complex paragraph but without full comprehension. By the end of second grade, a child should reach Level 3: the ability to read and fully comprehend a more complex paragraph. The levels of performance established for the ECE are inclusive. That is, the students that reached Level 3 are highly likely to be able to perform all reading tasks of this level plus those corresponding to the previous levels. The definition of meaningful performance levels can only be done by transforming scores using the Rasch model or a similar approach (MINEDU-UMC, 2007).

c. Oral measures of reading fluency

More than 25 years of research support the use of oral reading assessments as highly reliable and valid measures of general reading achievement, including comprehension, for most students.[7] Oral reading assessments are designed to be inexpensive, quick measures to regularly monitor the development of pre-reading and early reading skills. They are mostly used by teachers to track individual student progress in reading, and identify students that require additional support.

All oral assessments require one-to-one administration. They can be done by teachers themselves and take between 10 to 15 minutes per student (RTI International, 2008; University of Oregon, 2005). Known examples of oral assessments are the AIMSweb Standard Reading Assessment Passages (also available in Spanish); the Dynamic Indicators of Basic Early Literacy Skills (DIBELS), the Developmental Reading Assessment K-3 - 2nd Edition (DRA-2, K-3), the Early Reading Diagnostic Assessment (ERDA), and the Early-Grade Reading Assessment (EGRA). Most of them have been field tested only in the USA, but EGRA has been piloted in twelve languages and six countries, including Peru (RTI International, 2008).

All these tests measure oral reading fluency (hereafter, fluency or reading fluency) the same way (by counting the number of words read correctly from a passage in one minute), and all include

7 See Shinn & Shinn (2002) for a useful summary of validity studies since 1981 and Hosp (2001) for a comprehensive review of correlational studies as early as 1981.

measures of pre-reading skills like phonemic awareness and phonics. For example, EGRA includes timed, one-minute assessments of letter recognition, nonsense and familiar words, and paragraph reading. Additional (untimed) segments include comprehension, relationship to print, and dictation (RTI International, 2008). The DIBELS include two measures of Phonological Awareness: Initial Sounds Fluency (ISF), and Phonemic Segmentation Fluency (PSF). ISF assesses a child's skill to identify and produce the initial sound of a given word, while PSF assesses a child's skill to produce the individual sounds within a given word. Alphabetic understanding is evaluated with the Nonsense Word Fluency (NWF) (University of Oregon, 2005). Oral tests do differ on how they measure comprehension of the passage read. For instance, EGRA uses factual (literal) question-answering, while DIBELS uses a Retell Fluency scale (Hudson, Lane, & Pullen, 2005).

In most cases, concurrent and predictive validity is assessed using a variety of other standardized reading assessments such as the Iowa Test of Basic Skills- ITBS, California Achievement Test - CAT, Stanford Achievement Test - SAT, Colorado Student Assessment Program - CSAP, etc. (The Psychological Corporation, 2003; Torgesen, Wagner, & Rashotte, 1999; University of Oregon, 2005; Wiederholt & Bryant, 2001; Williams, 2001).

d. Correlations between reading fluency and "pencil-and-paper" comprehension measures

International research has demonstrated once again that measures of fluency and written comprehension are correlated (Abadzi, 2006; Crouch, 2006; Samuels, 2002). Deno et al (1980) found significant correlations between written comprehension and oral reading in three separate studies with regular students and students with reading disabilities (resource students). The first study was done with a sample of 18 regular students and 15 resource students, correlations between oral reading and comprehension correct rate of response were .67 for regular students and .82 for resource students. The second study was done with 27 regular students and 18 with reading disabilities, correlations between word meaning and oral reading were .49 to .57 for the regular students and 0.66 to 0.71 for resource students, depending on whether oral reading lasted 30 or 60 seconds. The third study was done with 43 regular students and 23 resource students. For regular students, the correlation between oral reading and SAT literal comprehension was .73, and .74 with inferential comprehension; for resource students, correlations were .7 and .67 respectively.

Fuchs, Fuchs, and Maxwell (1988, cited in Fuchs, Fuchs, Hosp, & Jenkins, 2001), found a correlation of .91 between reading fluency and comprehension in the SAT in a sample of 70 students with reading disabilities. Reading fluency also correlated highly with other measures of reading abilities. The authors concluded that oral reading fluency could be considered as an indicator of overall reading competence.

Shaw and Shaw (2002), found similar results for the relationship between oral fluency (measured with the DIBELS) and written comprehension (measured with the reading scale of the CSAP). In a sample of 52 students, the authors found significant, positive and strong correlations between scores in the CSAP, taken in the spring, and oral fluency measures taken in the fall, winter and spring. Coefficients ranged between .8 and .93, which led the authors to conclude that DIBELS is a strong predictor of reading performance in the CSAP (Shaw & Shaw, 2002).

Likewise, Wood (2005) found significant correlations between oral reading fluency and curriculum-based tests. Results reveal high percentages of variance in the latter accounted for fluency in first, third, and eighth grades, ranging from 89 percent to 69 percent (Wood, Hill, Meyer, & Flowers, 2005). He also found that oral reading fluency adds explanatory value even when other factors are held constant.

With correlation coefficients ranging from .49 to .91, one wonders what factors could explain this variation. A review of 21 empirical studies that report in total 33 correlations between text fluency and reading comprehension calculated that the mean correlation was .69, with a standard deviation of .11 (see summary table in Annex 7.2). The author found statistically significant differences based on: (1) the difficulty of the passage; (2) the length of passage; (3) the type of comprehension measure used; (4) the student's grade; (5) whether students have learning disability or not; and (6) sample size (Hosp, 2001):

- First, the mean correlation for studies using below grade level passages (.73) was higher than those using passages at grade level (.68). This difference was significant.[8] Hosp (2001) attributes this difference to the possibility of students at upper grade levels reaching a ceiling, which would result in a non normal distribution.

- Second, the highest mean correlation (.91) occurred when passages of 100 words were used, and the authors indicated that scores were prorated when necessary, longer (up to 400 words) or shorter (90 words) passages yielded lower correlations. The author attributes this to the fact that students may complete a shorter passage before one minute, requiring examiners to prorate scores; however, this does not explain why correlations were lower in longer passages (Hosp, 2001).

- Third, the highest mean correlation for a criterion measure was reported for the maze or cloze (.78). This was significantly different from the standardized, group-administered tests (.69)[9]; and the Kaufman Test of Educational Achievement- Brief (.41).[10] This could be because the word-reading tasks and the maze measure are markedly similar. Also, multiple-choice questions are more likely to induce guessing. Yet, although the standardized, group-administered tasks were similar, the correlations for each test ranged from .51 to .85. Most notably, among these, the California Achievement Test resulted in the lowest figures. Hosp noted that by eliminating the four correlations reported in one study using the CAT, the mean correlation for group-administered criterion tests increased from .69 to .74.

- Fourth, studies that compared results by grades found that correlations are higher in lower grades (1 through 4), and then dropped off after fifth grade, but cross-study comparisons were not done (Hosp, 2001). The study in Junin also found lower correlations in sixth grade than in second and third grades (Banco Mundial, 2008).

- Fifth, correlations based on special education students were significantly higher (.77) from those based on general education students (.66)[11], and from those based on students at-risk (.66).[12] Studies that combined students in both groups reported a significantly higher mean correlation (.71) than studies only including students in general education (.66).[13] However, only two included used students with disabilities, and only two studies

---

[8] $Z(1, 1571) = 1.91$, $p < .05$.
[9] $Z(1, 1656) = 2.66$, $p < .01$
[10] $Z(1, 267) = 3.00$, $p < .001$
[11] $Z(1, 1338) = 2.48$, $p < .01$
[12] $Z(1, 264) = 1.83$, $p < .05$
[13] $Z(1,2081) = 2.12$, $p < .05$

included students at-risk, which weakens the conclusions. Fuchs, Fuchs, Hosp, & Jenkins (2001) noted, regarding their own research, that for students with reading disabilities "individual differences in word reading processes are likely to have a stronger effect on comprehension outcomes than among more skilled readers" (p.245).

- And sixth, there was a significant difference between the mean correlation reported for studies with a sample size below 50 reported larger correlations (.73) than those with a sample size between 50 and 99 (.65), and differences were statistically significant.[14] At the same time, the mean correlation provided by studies using a sample size below 50 is less reliable. For correlational research, a minimum of 30 subjects is considered necessary. A small sample size may affect the reliability of the correlation, meaning the chances of obtaining a similar score with the same population decreases with sample size (Hosp, 2001).

e. Choosing between oral and written reading tests

Different measures of reading skills serve different purposes, thus, the choice of test depends on its intended use. Most of the current literature focuses on the pedagogical purposes of measurement, distinguishing between diagnostics, screening, progress monitoring, and outcome evaluation (NRP, 2000). A *diagnostic* assessment should help teachers plan instruction by providing in-depth information about students' skills and instructional needs. The *screening* assessment is designed to identify students who may be at high risk of academic failure and in need of additional reading instruction. Reading assessment for *progress monitoring* should help determine if students are making adequate progress or need more intervention to achieve grade-level reading outcomes. And *outcome* assessments offer information on the effectiveness of the reading program in relation to established performance levels (Kame'enui & Simmons, 1990; Kame'enui & Simmons, 2000; The Access Center, 2005).

Reading assessment is key for decision making at the policy level as well, specifically regarding how to allocate resources more efficiently to improve quality and reduce inequality in education. Policymakers and administrators need information on the overall performance of the education system (*diagnosis*), the existing gaps between schools and student subpopulations (*screening*), the progress being made by schools or school districts with the resources available (*monitoring*), and the impact of government-mandated and alternative programs (*outcome evaluation*), so that they can make decisions. In this sense, assessment for policy purposes must focus on schools or school jurisdictions as the units of comparison, instead of individual students, keeping an eye on vulnerable groups in particular.

With clear objectives and intended audience in mind, choosing the most appropriate measure then depends on test characteristics. Here we compare two general approaches:[15] (1) Oral tests, which are usually criterion-referenced measures (CRMs),[16] individually administered several times a

14 $Z(1, 1357) = 2.66$, $p < .01$

15 The comparison between these two approaches is meaningful as long as it is limited to formal, valid, reliable tests (which are available for both approaches).

16 Criterion-referenced measures (CRMs) evaluate how well a student performs against an objective, standard or criterion (in this case, a given number of words correctly read per minute) rather than another student. Students are informed of the expected standard and are taught to succeed on related outcome measures. Hence the "bell curve" is skewed heavily to the right (ultimately all students are expected to succeed). In this sense, CRMs help eliminate competition and may improve cooperation.

year; and (2) pencil-and-paper tests, typically standardized, norm-referenced measures (NRMs)[17] administered to groups in large scale once a year or every number of years.

For ***diagnostic*** purposes in the classroom, oral measures are most appropriate because teachers can use them whenever they need information to make a decision about selecting instructional materials or methods, forming instructional groups, providing remediation and enrichment, and assessing the need for individualizing instruction. Standardized national tests, on the other hand, will not be able to produce and report that information early enough in the school year so that teachers have a chance to incorporate the results in the decisions they make (Iowa Testing Programs, 2005). At the policy level, any diagnosis of how the education system is performing requires large samples in order to be representative, for which pencil-and-paper tests are most appropriate. However, these tests have more limitations in assessing pre-reading and early reading skills, compared to oral testing. In developing countries –especially in poor, rural settings–, assessing pre-reading skills in the first grades is key to obtaining an accurate diagnosis of early problems for policy purposes, as the lower end of the performance curve tends to be longer. Oral tests can serve this purpose and be administered in large enough samples, as it was done in the Junin region, in Peru, by The World Bank (Banco Mundial, 2008), bringing attention to the particular issues that are faced by specific subsets of at risk groups –such as indigenous students or rural single teacher schools.

For ***screening*** students at risk of academic failure, individually administered oral measures are more dependable than group assessments because the teacher can command the mental engagement of the student, to some extent, during the testing process (McKenna & Stahl, 2003). Moreover, standardized pencil-and-paper tests do not discriminate well until enough progress has been made by children (Hess, 2006). Most written tests assume students are already competent in phonological awareness, phonics and decoding; are comfortable using pencils; and can understand the concept of choosing an answer that is already being provided in the question (most students are used to answer open-ended questions in class, similar to those asked in oral tests). In general, CRMs provide a finer measure of beginning skills, as the distribution curve is skewed to the right and the left tail is longer. In particular, tests like DIBELS or EGRA are designed to assess the most basic foundation skills for literacy acquisition in grades 1–3, including pre-reading skills such as phonemic awareness, alphabet awareness or listening comprehension (RTI International, 2008).

At the policy level, CRMs provide more useful information than NRMs for determining schools at risk and require additional support. At risk schools should be selected based on how far are they from reaching certain educational goals, especially if they fell off track, because the important thing is to bring them back on track. Their position relative to other schools is less meaningful, as they could be last but already meeting curricular goals. On the other hand, NRMs are appropriate to identify positive outliers, that is, effective schools: those that achieve more with similar resources. Lessons can be drawn from the best schools, even if they do not reach the standard (sometimes standards are set too high), because the important thing is to find out why are they doing better than most.

For ***monitoring*** students' progress in early skills, oral tests provide more and timely information, because teachers can use them with the necessary frequency during the school year and the

[17] Norm-Referenced Measures (NRMs) are designed to determine individual performance in comparison to others, using standardized population norms. In these tests, items produce great variance in scores, perhaps with less than 50% scoring correctly. In this sense, distribution of scores usually follows a normal curve. Items cover a broad range of content and often represent a mismatch between what is taught locally and what is mandated in the general curriculum.

individual administration gives them detailed information on what particular areas need to be reinforced for each student. This requires comparable alternate forms of testing sensitive to change in performance over short periods of time. Also, when reporting to parents, teachers need some idea of what constitutes “grade-level performance” using references in the form of grade-level benchmarks (criterion-based), rather than using statistical percentiles (Hess, 2006). For example, DIBELS provides teachers with benchmark scores that students need to meet at specific times during each year and suggests instructional interventions that are needed by students who fail to meet benchmark scores (University of Oregon, 2005). In contrast, monitoring progress of schools for policy-related purposes can be done annually or multiannually (Barksdale-Ladd, 2000), and in the case universal coverage is recommended (so each school receives feedback). In Peru, the census assessment (ECE) meets the requirements to accomplish this goal: it is universal, annual, and there are no high stakes that could complicate even further the testing process (although its end-of-the-year application can lead teachers to believe that there are high stakes attached).

Last of all, for ***evaluating outcomes*** of an instructional program, segment or intervention at the classroom level both types of tests are appropriate, as long as they are administered to all –or a representative sample of– students and are CRMs. CRMs, are most appropriate for assessing what concepts and skills students have learned from a segment of instruction. Most CRMs are individual oral assessments, but some standardized written tests can also become CRMs by, for example, transforming scores with models such as Rasch, a standardized linear logit transformation with given weights[18] (Bond & Fox, 2007; Wright & Stone, 1979).

NRMs, such as pencil-and-paper tests that use raw scores or percentages of right answers are not appropriate for determining the effectiveness of educational programs because: (1) there is not, a priori, any value that can be considered as satisfactory or unsatisfactory performance; (2) raw scores do not indicate what students know or not know; (3) they do not take into account the difficulty of the questions, as all questions have the same value; (4) they cannot refer in any way to the contents of the questions; (5) they do not give information on the relative importance of the questions not answered correctly, or how many subjects did not answer them; and (6) they do not allow for comparisons between different tests, that is, it is incorrect to say that a mean of 51 percent of right answers in one test is better than a mean of 45 percent in another test[19] (Bazán & Millones, 2002).

The Rasch model can overcome all these limitations, and provide information that has more pedagogical value and potential for comparison over time because it can cross-reference the results in each item with the curricular standards that constituted the framework for the elaboration of the test. It also allows the reporting of results in a transformed scale that can be compared to other scales. All of this enables the written test to become a criterion-referenced measure.

In sum, “teachers need information specific to the content and strategies they are teaching, and they can best get that information through assessments built around their daily educational tasks”

[18] The Rasch model is particularly used in psychometrics –the field concerned with the theory and technique of psychological and educational measurement– for analyzing data from assessments of abilities, attitudes, and personality trait

[19] It is possible to address some of these limitations by doing an item-by-item analysis with raw scores. The result would be to place each child in a percentile relative to the sample for each single question. Thus, for example, in question 1 (reading a word), child “i” fails while 95 percent of children answer correctly, this only tells us that he is in the bottom 5 percent in this item. This type of analysis is even more cumbersome than using Rasch scores, and does not provide an adequate integrated picture.

(International Reading Association, 1999, p. 4). For most teaching purposes (all but outcome evaluation), the best alternative are low-cost, easy to administer tests. Such tests produce timely information to plan lessons and strategies based on the skill level of students, to track reading improvement for individual students, to identify children who are not demonstrating adequate progress, and to improve instruction strategies on time so that at-risk learners do not fall further behind. But in any case, teachers must have training in the assessment strategies they use and feel comfortable with their implementation. Most importantly, any assessment must be instructionally relevant, focused on essential skills, and culturally and linguistically appropriate (Skiba, Simmons, Ritter, Kohler, & Wu, 2003).

In contrast, for most policy-related purposes, group-administered pencil-and-paper tests, such as the one being implemented in Peru by the MINEDU, are far more efficient (McKenna & Stahl, 2003), because they allow greater coverage of the student population under standard testing conditions, and thus offer results that are comparable across schools and regions. Standardized norm-referenced or criterion-referenced tests provide a broad picture of achievement on certain kinds of tasks that are valid indicators of quality of instruction (International Reading Association, 1999). Written tests can also be purposely designed to inform policy decisions. For instance, the Programme of International Student Assessment (PISA) is policy-oriented, with design and reporting methods determined by the need of governments to draw policy lessons (OECD, 2007).

Based on the four purposes of evaluation discussed before, Table 2 offers a quick summary of the best uses for oral and pencil-and-paper tests in early reading evaluation when it comes to teaching decisions –such as planning lessons, selecting materials, providing remediation assistance, and so on– and policy decisions –such as assigning resources, designing in-service teacher training programs, implementing targeted programs, revising the core curriculum, designing textbooks, and so on.

Table 2. Summary of best use for oral and standardized pencil-and-paper tests

Purpose	Test format (administration)	Teaching decisions	Policy decisions
Diagnosis	**Oral (individual)**	**YES.** Produce timely information on individual students that helps electing instructional materials or methods, forming instructional groups, providing remediation and enrichment, and assessing the need for individualizing instruction.	**NO.** They are costly and time-consuming to implement on a large scale. **BUT** can produce valuable qualitative information to triangulate with results from larger samples, particularly when diagnosis on early reading skills is needed.
	Pencil-and-paper (group)	**NO.** Do not produce results early enough in the school year to inform pedagogical and administrative decisions at the school level.	**YES.** Allow for large-scale testing, produce comparable results. This can be done by comparing schools relative to other schools (NRM) or relative to a goal (CRM).
Screening	**Oral (individual)**	**YES.** More dependable, discriminate better among low-performers and in early grades. CRMs are most appropriate because they discriminate better among poor-performers than NRMs.	**NO.** Not dependable. Too costly to implement for all schools. **BUT** could be implemented in schools that have been already screened, for a better understanding of needs.

Table 2. Summary of best use for oral and standardized pencil-and-paper tests

Purpose	Test format (administration)	Teaching decisions	Policy decisions
	Pencil-and-paper (group)	**NO.** Assume students are already competent in phonological awareness, phonics and decoding; are comfortable using pencils; and understand the mechanics of multiple-choice tests.	**YES.** Can be used to screen for schools performing poorly to develop targeted programs (CRMs are better for this), as well as for effective schools, to draw lessons to improve policy frameworks (NRMs are better for this).
Monitoring progress	**Oral (individual)**	**YES.** Provide more and timely information. CRMs give teachers and parents an idea of what constitutes "grade-level performance" with benchmarks.	**NO.** Too costly to implement for all schools.
	Pencil-and-paper (group)	**NO.** It cannot be done with enough frequency to provide feedback.	**YES.** Allows universal coverage (recommended).
Outcome evaluation	**Oral (individual)**	**YES.** As long as it is CRM and administered to all students in the class or a representative sample.	**NO.** Too costly to implement for all schools.
	Pencil-and-paper (group)	**YES.** As long as it is CRM and administered to all students in the class or a representative sample.	**YES.** NRMs are most appropriate when the purpose is to compare outcomes across schools or jurisdictions. CRMs are most appropriate for determining the effectiveness of educational programs.

Source: Prepared by the authors. The table reflects the assessment of the authors based on their review of literature and field experience with both tests

Now, beyond the four purposes commonly identified in the literature, evaluation of student performance is also an important accountability tool in education: schools are given an amount of resources and then they are held accountable for producing good results; the same is true for teachers, principals, local and national education officials. Because students enter first grade with different levels of exposure to language and print depending on their socioeconomic background, accountability against fixed standards in first and second grades is not appropriate. Instead, teachers, schools, education officials should be held accountable for producing progress in relation to the starting point. That is why diagnosis, monitoring and outcome evaluations are all needed for accountability.

For downward accountability from the school to the students, the parents, the community, oral reading tests that are criterion referenced can be used by parents and school boards to verify that students in the early grades are learning the basic skills. Oral tests are easy to apply by lay people and relatively low cost. On the other hand, pencil-and-paper tests are more appropriate for any type of systemic accountability, including upward reporting. End-of-the-year testing –such as the one carried out in Peru by the UMC– allows or encourages the use of scores to evaluate schools, administrators, and individual teachers because the scores are obtained very close to the end of a major instructional segment (the school year).

Having said that, using early grade reading measures for high-stakes accountability –e.g. to promote or fire personnel, award grants to schools, etc.– is not recommended. Some authors argue that before the age of 8, standardized pencil-and-paper achievement measures are not sufficiently accurate to be used for high-stakes decisions about individual children, teachers and schools, and advocate for beginning-of-the-year evaluations instead, precisely to discourage perception of teachers that the scores might be damaging to them professionally (International Reading Association, 1999; Iowa Testing Programs, 2005).

f. Keeping in mind tests sometimes behave differently with certain groups

Academic performance tests (or items) do not always measure skills the same way for different groups of students. Different methods for analyzing differential item functioning (DIF) have been in used since the mid 1980s to determine whether a test is being "unfair" to certain populations (see Dorans & Kulick, 1986; Holland & Thayer, 1988; Shealy & Stout, 1993; Swaminathan & Rogers, 1990). DIF occurs when different groups of examinees with the same level of proficiency in a domain have different expected performance on an item; in other words, students with the same ability find it differentially difficult (Mapuranga, Dorans, & Middleton, 2008). Thus, DIF analysis is an important step when evaluating tests for fairness and equity. In DIF analysis the sample is usually divided into two subgroups, but additional subgroups are possible. The reference group provides a baseline for performance (e.g. non indigenous or male) and the focal group is the focus of fairness concerns (e.g. indigenous, black or female). An external reference measure (matching variable) is used to establish score equivalence between the two groups so that difference in performance in a given item can be attributed to DIF and not to other factors. The matching variable can either be an observed or unobserved score and either a univariate or multivariate variable.

Studies have also shown that contextual factors (e.g., language and cultural characteristics) can have an impact on DIF analyses, particularly for non-native speakers. For example, it was found that when Hispanics took the SAT Reasoning Test (formerly the SAT Verbal test), items containing specific linguistic features (e.g., true cognates, false cognates, homographs) and items that are of special cultural interest exhibited DIF. When native language speakers are tested in their native language, DIF occurs on some items between them and non-native speakers due to language familiarity learned outside the classroom: some examinees use the language being tested as an "academic language" while others use the language as a "home language" (Mapuranga, Dorans, & Middleton, 2008). Familiarity of test language goes beyond linguistic references to also include the phrasing and formatting of items (e.g. multiple-choice) and their content, which could be more familiar to certain groups of students than for others. When the item (or test) is administered in different languages, differences in language difficulty and meaning are the most likely explanations (Emenogu & Childs, 2005).

Another possible explanation for differential behavior of items and tests is provided by the concept of stereotype threat, which is the fear –conscious or unconscious– that one's behavior will confirm an existing negative stereotype of a group with which one identifies. This fear has been proven to lead to an impairment of performance in tests both in racial minorities and women. When a fixed biological characteristic such as race or gender is emphasized before a given test, test performance is negatively affected among members of groups for which culturally-shared stereotypes suggest poor performance (Steele & Aronson, 1995).[20] Low performance can also derive from a self-protecting reaction, by which one does not give it the best try in mastering the test because assumes that, no matter what, the stereotype cannot be changed.

[20] The authors conducted several studies. Studies 1 and 2 varied the stereotype vulnerability of Black participants taking a difficult verbal test by varying whether or not their performance was ostensibly diagnostic of ability, and thus, whether or not they were at risk of fulfilling the racial stereotype about their intellectual ability. Reflecting the pressure of this vulnerability, Blacks underperformed in relation to Whites in the ability-diagnostic condition but not in the nondiagnostic condition (with Scholastic Aptitude Tests controlled). Study 3 validated that ability-diagnosticity cognitively activated the racial stereotype in these participants and motivated them not to conform to it, or to be judged by it. Study 4 showed that mere salience of the stereotype could impair Blacks' performance even when the test was not ability diagnostic (Steele & Aronson, 1995).

2.3. Linking reading development standards with measures: the case of Peru

The Peruvian government sees universal testing as the best way to provide feedback on outcomes to the education system and communities, in line with its policy of increasing local participation in monitoring service delivery. In order to do this, Crouch (2005) recommends that results from the student performance assessments be based on curricular contents, and measured against the standards set for each grade. Having clear standards would also allow teachers and the education community in general to be accountable for results, provided that they are given the necessary resources to produce them.

Standards should be well-written and reasonable, but in many cases –in the spirit of making them comprehensive and raising the bar– they end up including too much, being too vague, being too difficult, and undermining local curriculum and instruction. If the standards are flawed or limited, tests based on them also will be. Even if standards are of high quality, it is important to know how well a particular test actually matches the standards (FairTest, 2007).

Today, Peru has a diversified approach to basic education curriculum, leaving considerable room for regional, provincial, local, and school authorities and teachers to incorporate locally relevant contents and activities to the core curriculum (MINEDU, 2005). In the early grades, however, diversity of curriculum is not so much of an issue when it comes to measurement because there are a set of basic skills such as reading and mathematical reasoning that are the foundation for future school learning and common ground for all schools.

The national curriculum defines a set of expected learning outcomes in reading for each grade, but all these outcomes refer only to passage comprehension (MINEDU, 2005). No standards are set for pre-reading (e.g. phonics, phonemic awareness) or early reading (e.g. letter recognition, pseudoword decoding or word reading) skills. In other countries there are preliminary goals set for the early grades, which are very specific. This is the case of both developed (e.g. Finland, the US, Chile) and developing countries (e.g. South Africa, Kenya). For example, Delaware's *English Language Arts Grade-Level Expectations* establish reading standards that include decoding skills and other word recognition strategies for grades K-12, as well as vocabulary development. Kenya's Primary Education Syllabus includes standards for pre-reading skills such as distinguishing different sounds, developing oral vocabulary, identifying differences in shapes, sizes and colors. Early reading skills include naming letters and differentiating the sounds of syllables (Kenya Institute of Education, 2002). The Learning Outcome 6 of the South African national curriculum states that "[t]he learner will know and be able to use the sounds, words and grammar of the language to create and interpret texts. Learners will begin to explore how language works […] they will learn and use terms like 'sound', 'syllable', 'word', 'sentence' and 'full stop'. This will enable them to discuss and think about such things as how words are spelled and how simple sentences are structured and punctuated." (Department of Education, 2002, p. 12).

In any case, the UMC designs the national evaluation for second grade based on the curricular standards, which are succeptible of interpretation (MINEDU-UMC, 2007; MINEDU, 2005). From what we can see, second grade Peruvian students are expected to be able to: (1) recognize characters, main events, materials, ingredients and instructions in narrative and descriptive texts; (2) infer facts and ideas based on hints from the text and previous experiences; and (3) form and express an opinion about the facts and ideas of the text. In *third grade*, the students must strengthen those abilities and also become familiar with informative texts, being able to: (4) recognize data, themes and places; (5) infer important information and identify the subject from information explicit in the text and previous experiences. In *fourth grade*, students should also be able to: (6) understand more complex texts; (7) recognize the type of text they are reading; (8)

differentiate between the main and secondary ideas in the text; (9) recognize themes and procedures; and (10) infer the purpose for which the text was written. In *fifth grade*, students must be able to, on top of everything else: (11) read argumentative texts; (12) use the details of the text to make inferences; and (13) express a critical opinion about the content and form of the text. Finally, in *sixth grade* students should learn to: (14) infer ideas from information that is only implicitly presented in the text (MED, 2005).[21]

Table 3. Match between items in the ECE test and Curricular Standards

Passage	Question	Item*	Corresponding Curricular Standard
Words	1	1	First grade
	2	2	
	3	2	
	4	3	
	5	3	
Sentences	6	4	
	7	4	
	8	5	
	9	5	
Narrative (short story)	10	6	(1) recognize characters, main events, materials, ingredients and instructions in narrative and descriptive texts
	11	7	
	12	8	(2) infer facts and ideas based on hints from the text and previous experiences
	13	9	
	14	10	
	15	11	(5) infer important information and identify the subject from information explicit in the text and previous experiences (3^{rd} G)
Informative (poster)	16	12	(4) recognize data, themes and places (3^{rd} G)
	17	13	
	18	14	(10) infer the purpose for which the text was written (4^{th} G)
Descriptive (longer text)	19	15	(1) recognize characters, main events, materials, ingredients and instructions in narrative and descriptive texts
	20	16	(1) recognize characters, main events, materials, ingredients and instructions in narrative and descriptive texts
	21	17	(4) recognize data, themes and places (3^{rd} G)
	22	18	(5) infer important information and identify the subject from information explicit in the text and previous experiences (3^{rd} G)
	23	19	(10) infer the purpose for which the text was written (4^{th} G)
Second grade skills not measured			(3) form and express an opinion about the facts and ideas of the text
Post second grade skills not measured in the test			(6) understand more complex texts (7) recognize the type of text they are reading (8) differentiate between the main and secondary ideas in the text (9) recognize themes and procedures (11) read argumentative texts (12) use the details of the text to make inferences (13) express a critical opinion about the content and form of the text (14) infer ideas from information that is only implicitly presented in the text

* For scoring, some questions were merged because they were interdependent, thus changing the number of items to 19.
Source: Prepared by the authors, based on the ECE test and the National Curricular Design (MINEDU, 2005)

When designing a test, including some questions to measure below and above grade-level performance is useful to expand the distribution of scores and screen students who may have fallen behind or who are getting ahead. In the case of the ECE test for 2006 (used in this study) the first nine questions (out of 23) of the test measured what could be classified as first-grade skills, while

[21] Numbered by us for clarity purposes.

other seven questions measured third and fourth grade level skills according to curricular objectives. This left seven questions to measure grade-level abilities (see Table 3), which may be not enough to adequately discriminate performance of second graders. In addition, there is one skill (critical comprehension) expected for second grade that is not included in the test, as it is much harder to measure in standardized, multiple-choice tests like the ECE.

Curricular objectives in terms of skills to be mastered are not standards in and of themselves, but they serve as guidelines to set relevant standards. Standards can be defined in two ways, relative to a population norm (Norm-Referenced Measures, NRM) or relative to a fixed criterion (Criterion-Referenced Measures, CRM). For standardized tests, CRMs make more sense than do NRMs, because the purpose is to know how many students have mastered the skills expected for their grade.

Some CRMs, however, are not based on a specific curriculum, but on a more general idea of what students might be taught. A recent variation of criterion-referenced testing is "standards-referenced testing" or "standards based assessment" (SBA), which differ from "curriculum based measurement" (CBM) in that the emphasis of SBA is on clearly defined standards (which derive from curricular frameworks), while CBM does not necessarily rely on standards.

Many states and districts in the US as well as in many developed and developing countries (such as the UK, Chile, South Africa, Kenya) have adopted curriculum frameworks that describe what students should know and be able to do in different subjects at various grade levels, similar to what the Peruvian curriculum framework offers (MINEDU, 2005). But these often have performance standards (which may be derived from the curriculum but might not be part of the highest-order curricular materials) that define how much of the content standards students should know to reach the "basic," "proficient" or "advanced" level in the subject area. Their tests are designed based on the standards, and the results are reported in terms of these levels.

In Peru, the UMC uses the Rasch model (Bond & Fox, 2007; Wright & Stone, 1979) to transform scores of the standardized tests, thereby making it possible to report results of an otherwise normative test in reference to a set of criteria and by performance levels (MINEDU-UMC, 2004). That is why it was possible to provide feedback to students, teachers and parents regarding the level of performance that the child mastered in the ECE test (MINEDU-UMC, 2007). However, it is still not clear how the standards supporting the definition of these levels were set or by whom.

In CRMs, the passing or "cut-off" score (criterion) is typically decided by a committee of experts (or by the teacher if the test is administered in the classroom). In both cases, deciding the passing score is subjective, even when using empirical research to back it up, and ultimately represent human judgment. Often, the criterion set does not take into account the factors influencing minority students' performance, and could affect their pass rates considerably (FairTest, 2007). Similarly, the norms used in NRMs are developed based on standardizations taken from a general population. When used for important decisions regarding student placement and advancement (high-stakes testing), for example, they often put minority students at a disadvantage.

In this sense, Fuchs and Fuchs (2002) point out that in order to bolster the meaningfulness and usefulness of a CBM as an education accountability tool it is important to take into account how CBMs incorporate students that do not fall under the normative range, such as students with learning disabilities. Since in Peru reading performance levels are currently quite low (Crouch, 2005), the issues raised by the authors may be of relevance when defining standards and designing tests.

First, national normative data –representative of all groups and regions– are required before a CBM can achieve status as a universally accepted and implemented accountability tool, because as such it will be used to formulate comparisons between states, districts, schools, and teachers about effectiveness of education programs. The authors suggest that normative data should: (a) provide information about typical performance level at each trimester of the school year for grades 1-6; (b) provide information about slopes at each grade for typically developing children as well as those with high-incidence disabilities; (c) systematically sample widely used reading programs; and (d) address reading and other basic skills where well developed CBM procedures are specified (Fuchs & Fuchs, 2002).

Second, Fuchs and Fuchs (2002) argue for the need to prevent unintended consequences. They recommend conducting studies that examine the effects of an outcomes-based accountability framework on the ambitiousness of learning goals, the quality of teachers' instructional plans, and the extent of student learning. Such studies should also identify how aggregating data by teachers, service delivery arrangements, instructional methods, curriculum packages, and types of disabilities affects decision making at the school, district, state, and national levels.

Third, it is necessary to expand the focus not only to the acquisition of basic skills in reading, mathematics, spelling, and written expression at the elementary grades, but also to more complex skills and knowledge application in later grades (Fuchs & Fuchs, 2002). As Crouch (2005) indicates, the objectives of quality secondary education could be reflected in a national graduation certification exam. Again, norms and graduation criteria should be set in a way that they do not discriminate against already vulnerable groups.

Fourth, Fuchs and Fuchs (2002) suggest that is necessary to study the methods for aggregations of data at the classroom (teacher), school, district, and regional levels, so that important information is not lost in the process: information on the effectiveness of different classrooms or schools, of contrasting service delivery arrangements, of various instructional methods or curriculum packages, and for different types of educational needs. This information is helpful to administrators and policymakers in specifying and improving education practice, but when merged into, for example, district averages, it gets lost. Related to this point, Crouch (2005) recommends to establish standards for process, such as effective instructional time per year or teacher recruitment standards, as well as for user information, with a top-down accountability process through which local education management offices (UGELs) are accountable to the schools they serve, and schools are accountable to the communities they serve.

As pointed before, standards must be specific both to grade-level and curricular objectives. Regarding early reading skills, benchmarks must approximate for the cognitive development of children. A World Bank study in Peru (Crouch, Abadzi, Echegaray, Pasco, & Sampe, 2005) determined that a child should be able to read correctly, out loud, an average of 30 words per minute by the end of the first grade, 60 words per minute by the end of second grade, 90 by the end of third grade and 110 by the end of fourth grade, and continue to progress in the following grades.

These standards are below the benchmarks used in other countries. For instance, standards for fluency in connected text in the United States range from 40 to 60 words per minute by the end of first grade, 82 to 94 by the end of second grade, 100 to 114 by the end of third grade, 105 to 118 in fourth grade, and 118 to 128 in fifth grade (Hudson, Lane, & Pullen, 2005). Oral Reading Fluency norms fall within these parameters (Hasbrouck & Tindal, 2006). Suggested standards in Spanish-speaking countries are also higher than those recommended for second graders in Peru, for example: 70 words per minute in Chile (Barrientos, no date), 75 words per minute in Spain (Equipo de Orientación Educativa de Marbella 2003; cited in Abadzi, 2006).

Similar benchmarks are needed for reading comprehension, which, in this case, should reflect curricular objectives. Currently, the large gap between the level of achievement required on the national assessments and the actual performance of the students in Peru (Banco Mundial, 2008), may prevent curricular goals –or standards derived from them– from being reasonably challenging and leading to stronger learning outcomes. When receiving reports on the results of their students, teachers may perceive that national standards are out of their reach and therefore would not take them as serious parameters to guide instruction. It is certainly discouraging, especially if standards are so high that progress made from one year to the next is not perceived because results are still far below the expected level.

Besides the definition of standards, it is also important to foster an accountability system that provides a more proximal and sensitive framework for cataloging student learning (Fuchs & Fuchs, 2002), so that results can be shared with the community in general. If parents do not know what the expected reading performance is for each grade, they will be likely to accept whatever reading level their children are able to produce, as they seemingly improve over time (Banco Mundial, 2008).

2.4. Research in context

After reviewing the relevant literature and current research, there are a few points that require more research and discussion. Two in particular caught our attention. First, given the complexity of the cognitive processes involved in reading, could there be more to the relationship between reading fluency and comprehension measures than what pairwise correlations have told us so far? Second, based on the assessment experience in Peru so far, in particular with regards to the first universal reading test (the ECE 2006), what lessons can be drawn for improving future measures and for using them more effectively in guiding policymaking and strengthening accountability in education?

This following sections aim to contribute to answer both questions by comparing results obtained by 475 students on two measures of reading abilities in Junin: an individually administered oral test, and the standardized, written test used for the ECE 2006. The specific research questions to address the paper's objectives are: How accurately do oral and written tests assess reading skills? What is the relationship between fluency and comprehension measures? And, based on the answers to those questions: How can reading tests be used to guide policymaking and strengthen accountability?

3. FIELDWORK METHODOLOGY

This section begins with a description of the participants in the study, including the sampling method, characteristics of the sample, and the calculation of an adjustment factor to correct sampling biases. Then, it describes the two reading tests used, the calculation of final scores in both tests, and the conditions of test administration.

3.1. Participants

The target population of this study is comprised of third grade students from public primary schools in the region of Junin. There are 1,684 public primary schools in Junin, with a total of 32,700 third grade students enrolled (see Annex 7.4.a). The Junin study used a multistage cluster sampling approach.[22] The first stage applied a probability-proportional-to-size sampling method to select the clusters, which in this case were the schools.[23] To ensure there would be enough students from each type of school to run separate analyses and comparisons by groups, all public primary schools of Junin were divided into three subgroups by type of school, according to the categories established by the Ministry of Education: full grade schools, multigrade with more than one teacher (hereafter "multigrade") and multigrade with only one teacher (hereafter "single-teacher").[24] Schools were selected randomly from each group. The next stages consisted of randomly selecting elements within the original cluster: first, one section of each relevant grade in each school, and then a sample of students in each selected section (see Annex 7.3 for details). The oral test was administered to the random sample of students in each grade (5 students per class in 2nd and 6th grades, and 10 in 3rd grade). The written test was administered to all students from the selected third-grade classroom present at the time.

In this paper, we use a sample of all 475 third-grade students who took both the oral and written tests for the comparative analysis between the two tests.[25] The schools in the sample vary in size and location to account for the diversity of the population. The smaller school has 12 students and the largest has 1713, ranging from 1 to 52 teachers. All nine UGELs (the Local Education Management Offices mapped to the provinces) were represented and the distribution across provinces in the sample is similar to that of the population (see Table 20 in Annex 7.4.b).

Table 4. Number of third grade students by type of school

		Population		Sample	
		Freq.	Pct.	Freq.	Pct.
Total		**32,700**	**100.00**	**475**	**100.00**
Type of school	Single-teacher	3,157	9.65%	104	21.89
	Multigrade	12,582	38.48%	164	34.53
	Full grade	16,961	51.87%	207	43.58
School location	Rural	16,595	50.75	307	64.63
	Urban	16,105	49.25	168	35.37

About 52 percent of students in the sample attend rural schools, and 60 percent go to full grade schools. The sample does not follow the distribution across types of schools found in the population. This bias, introduced with the sampling strategy, is corrected in the analyses by adjusting the sample according to the real probability of selection in the population with an

22 Cluster sampling is a sampling technique used when "natural" groupings are evident in a statistical population. In this technique, the total population is divided into these groups (or clusters) and a sample of the groups is selected. The clusters should be mutually exclusive and collectively exhaustive. In single-stage cluster sampling, all the elements from each of the selected clusters are used. In multi-stage cluster sampling, a random sampling technique is applied to the elements from each of the selected clusters. In this case, the authors opted for the latter, as using all sample elements in all the selected clusters was too costly and not necessary for the purpose of the study.

23 In this method, the probability of selecting a cluster in any given population varies conversely with the size of the cluster. That is, the probability of a school being selected in the sample is higher for schools with larger student populations and lower for those with fewer students.

24 About 10 percent of third grade students in Junin attend single-teacher schools, 38 percent attend multigrade schools and 52 percent attend full grade schools.

25 The psychometric analysis of the written test was done with a broader sample including all 1,001 students who took the written test (whether they took also the oral test or not).

adjustment factor as weight (see Annex 7.4.c). By increasing the representation of the underrepresented groups we ensure that the statistical analyses for the whole sample keep similar proportions to those found in the population in terms of the type of school that students attend.

In this sample, 48 percent are girls and 52 percent boys; two thirds have had access to some kind of pre-primary education. The age-range goes from 7 to 13 years old, but 74.2 percent are in or below the normative age for their grade (8 years old). The vast majority (85 percent) learned Spanish as their first language, 8.4 percent are native Ashaninka speakers, 5.7 percent are from Quechua origin, and less than one percent spoke another indigenous language (see Annex 7.4.b).

3.2. Instruments

Reading performance was assessed through an oral fluency test and the Ministry of Education's written comprehension test used in the Census Evaluation of 2006 (ECE). The instruments were administered in the first half of the 2007 academic year, between May and June.[26]

a. Oral test

The oral test measures both reading fluency and comprehension. It consists of a simple, short story –in this case of 60 words– which the child is asked to read aloud. It was administered one child at a time by a trained evaluator, without the presence of the teacher, parents or other students. The evaluator measured the time the child took to finish the passage, and registered the number of words read minus the words read incorrectly (if the child spontaneously corrects his or her mistake, it counts as a correct word). If the child gets stuck in one word, the evaluator can read the word to allow the child to continue, but it will not count as a correctly read word.

Students were given up to 4 minutes to finish the passage, unless they were not able to read the first words. If children are not given the opportunity to finish the passage they are less likely to respond correctly the comprehension questions (especially those inquiring about how the story ends), and thus the correlation between fluency and comprehension would be less valid. However, we thought it was unnecessary to grant unlimited time for students to finish reading because after a while it only creates pressure on the child.

This test uses a measure of *reading fluency* that corresponds to the words read correctly per minute by the student. The total score for the oral test was calculated by dividing the sixty seconds allowed for the test by the actual number of seconds that the student took to finish reading the passage and then multiplying the result by the number of words read. If the student did not finish the passage in 60 seconds, then fluency is equal to the number of words read correctly in 60 seconds. For instance, if a student finished reading the text in 30 seconds, the total score would be (60/30)*60=120 words read per minute; and if a student read only 30 words in 60 seconds, the score would be 30. Some students finish reading the text in less than 60 seconds but made mistakes. For example, one can finish the text in 30 seconds but make two mistakes, reading only 58 words correctly. In that case, the total score would be (58/30)*60=116.

Comprehension was evaluated with two direct measures: passage recall and question answering. After reading the passage, students were asked to narrate the story in their own words and points were assigned based on the amount and accuracy of the information recalled. One point was assigned for each of the four main events included in the narration (the dog went walking with its

[26] This is why in this study the ECE was applied to third graders even though it measures reading performance of second grade students at the end of the school year.

Box 1. Oral test passage: "Dogo"

Había un perrito gordo y peludo llamado Dogo. La familia con quien vivía lo quería mucho. Dogo era un perro obediente, cuidaba la casa y comía toda su comida. Un día salió de paseo con su dueño Lucas y se perdió. Lucas se puso triste, pero felizmente Dogo apareció al rato. Después se fueron los dos juntos a su casa.

There was a fat, furry little dog named Dogo. The family with whom he lived loved him very much. Dogo was an obedient dog, he watched the house and ate all his food. One day he went out on a walk with his master, Lucas and he got lost. Lucas was very sad, but fortunately Dogo showed up after a little while. Then, they went home together.

owner, it got lost, it was found, they went home), one point more if the events are narrated in order, another point if the student describes the dog (e.g. the dog was fat and furry), and one point more if he or she provides additional details. The student had to choose his/her favorite part of the story and then explain why he or she chose it, which could grant him/her up to two additional points. In total, this level of comprehension gets a maximum of nine points (*Recall Comprehension*).

The second measure was question answering. Students were asked five specific, literal questions such as "what was the dog's name?" For this level of comprehension (*Q&A Comprehension*) each correct response is worth one point, and the maximum score is five. When the student was not able to read more than one word of the text, comprehension scores were recorded as missing.[27]

b. Written Test (ECE)

The Ministry's written test included 23 multiple-choice questions (each with three options to choose from) regarding four passages, with different levels of difficulty. It was formulated based on the National Curricular Design (DCN), taking into account the learning achievements expected at the end of second grade in the area of integral communication, which for this grade focuses on reading comprehension. The test assesses both literal and inferential comprehension on three different types of texts: narrative, descriptive and informative. To answer each question, the student must choose among three alternatives.

For the written test, two sets of scores were calculated: raw and transformed (see Annex 7.5.a). The raw total score is the sum of correct answers, which adds up to 19.[28] The transformed scores were calculated using the Rasch model, which is what the MINEDU uses to analyze and report the results of its academic performance tests, and also the one employed in international student performance assessments such as LLECE.

27 In this sample, there were 43 students who did not read a word, and one student who read only one word correctly. Comprehension cannot be fairly evaluated if the student is not able to read the text. In such case, the student obtains a fluency score of 0 or 1, based on how many words he or she reads correctly, but the comprehension score is recorded as missing.

28 The test contains four pairs of connected questions, namely items 2 and 3, 4 and 5, 6 and 7, and 8 and 9. These items are two questions with the same three alternatives of answer, which means that answering one question reduces the chances of error on the other item. Each of these pairs is treated as one item, the same way the MINEDU did for ECE (assigning one point if at least one of the items in the pair is answered correctly). Thus, the test comprises 23 questions but only 19 items, and the score range goes from zero to 19. There is also the possibility of transforming those scores to percentages of right answers, which would take values from 0 to 100, but we preferred to keep them in their raw form for simplicity and because transforming values to percentages does not provide any additional information.

Using a broader sample of 1,001 third-grade students who took the ECE test,[29] we calculated three alternative scales of transformed scores with Rasch. The first scale was done with WinBUGS and includes all 1,001 cases. The second scale was calculated with Winstep, which automatically excludes the extreme cases and it is the method used by MINEDU (N=980). Finally, a common criticism to the use of multiple choice items in educational assessment is that there is no provision in the model for guessing, so a third scale was calculated using the Birnbaum's three-parameter logistic item response model to control for guessing (Baker & Kim, 2004; Birnbaum, 1968) but this model did not yield significant additional information. At the end, we decided to work with the first one because it provides information for all subjects and we found no need to control for guessing in this case (see Annex 7.5.a for more details).

After choosing the most appropriate Rasch scale, we analyzed the ECE test to evaluate which scores (raw or Rasch) would be better suited for comparison against the oral test. In theory, the Rasch model is more accurate, as it gives more information, has better psychometric properties and allows interpretation of the results from a pedagogical and curricular point of view. However, the benefits of using Rasch scores depend on the characteristics of the test and how it behaves in a given population. Unidimensionality is the most important premise for using the Rasch model. In this case, we find that the test is not unidimensional (Annex 7.5.c), which makes it unfit for using Rasch as a measurement model.[30] But even when the test cannot be fully adjusted to the Rasch model, transforming scores into a Rasch scale would yield useful information on the difficulty of items as well as provide more elements for interpretation based on curricular standards. For the purposes of this paper, such information is quite valuable.

Moreover, the Rasch scores correlate strongly and significantly with the raw scores ($r=0.995$, $p<.0001$), which means that, in this case, we can use almost indistinctively raw or Rasch scores when exploring the relationship between the ECE test and the oral fluency test. In terms of the psychometric properties of the written test (ECE) there are no significant differences between using raw or transformed scores. Both distributions of scores are non-normal, and have moderate internal reliability (Annex 7.5.c): in both cases the reliability coefficients for the sample of 1,001 students fall below the recommended threshold of 0.7, but they reach 0.74 for the sample of 475 students (which is the one finally used for the comparative analysis). Item analyses for both raw and Rasch scores indicate some problems with the construction of the test, specifically with regards to certain items (see Annex 7.5.b), but do not affect the behavior of the test as a whole.

Taking all these elements into account, we decided to use Rasch only for transforming scores and not as a measurement model (Bond & Fox, 2007; Wilson, 2004), which means that conclusions drawn from these analyses are suggestive but not definitive.

4. FINDINGS

In general, we find that the reading performance of third grade students in Junin is below the expected level both in terms of fluency and comprehension (Table 5). On average, it is estimated that a third grade student in Junin can read about 57 words in one minute, slightly below the

[29] To benefit from a larger dataset available, the calculation of Rasch scores was done using test results from all third-grade students who took the ECE test in the Junin study, totalling 1,001 (of which 475 also took the oral test). See Annex 7.4.b for a description of both samples and Annex 7.4.c for details on the adjustment factors.

[30] This issue deserves further discussion, but it is beyond the purposes of this paper.

standard for second grade of 60 words per minute. They obtain an average score of 3.45 over 9 in the open comprehension questions (Recall Comprehension), but do much better in the direct, literal comprehension questions (Q&A Comprehension), where they have an average of 4.3 over 5 with lower variation (SD=1). In fact, the majority of students (56.4 percent) were able to answer all five questions correctly.

On the written test, the average total raw score is 11.8 over 19, and the average Rasch score 45.5 over 75.[31] A student who finishes second grade is expected to be able to answer all questions correctly, but as noted before, several questions in the test are aimed at third or fourth grade skill level (see Table 3, in page 22). Based on probability calculations and difficulty levels of the items,[32] we estimate that a student with a Rasch score of 72 would be able to answer all second-grade questions correctly.

Table 5. Descriptive statistics of Fluency and Comprehension Variables (adjusted)

Test		Obs	Mean	SD	Min	Max
Oral Test	Fluency	475	57.41	29.59	0	157
	Recall Comprehension	431	3.45	2.83	0	9
	Q&A Comprehension	431	4.30	1.01	0	5
Written Test (Rasch Scores)		475	49.51	8.59	15.09	70.68

Next, we present the results of statistical analyses carried out to address the issues stated in the objectives of the paper, that is, to compare and evaluate how accurately the tests assess reading skills, to correlate scores of students in both tests, and to explore signs of discontinuity in the fluency-comprehension relationship.

4.1. How accurately do oral and written tests assess reading skills?

To compare the results in the oral and written tests we standardized all variables so that they are in the same scale, with mean of 0 and standard deviation of 1. Initial box plots suggest that there are no obvious differences in the distributions of scores and none of the distributions are normal (see Figure 16 in Annex 7.6).

Table 6. Assessment of Normality of Standardized Scores of Fluency and Comprehension Variables (Shapiro-Wilk W test)

Test	Variable	Obs	W	V	z	Prob>z
Oral test	Fluency	475	0.9876	3.978	3.311	0.00046
	Recall Comprehension	431	0.9588	12.139	5.961	0.00000
	Q&A Comprehension	431	0.9241	22.348	7.418	0.00000
Written Test (Rasch Scores)		475	0.9182	26.274	7.839	0.00000

31 The scale is set to range between 0 and 100 in theory, but 75 is the equivalent to 3 standard deviations above the mean, and so –although possible– no one is expected to score higher than 75.

32 Probability = $\frac{e^{(\theta-\beta)}}{1+e^{(\theta-\beta)}}$, where θ is the individual score, β is the item difficulty, and *e* is the exponential.

The lack of normality is confirmed by the Shapiro-Wilk W test[33] shown in Table 6. Next we evaluate how well these tests discriminate between high and low performers, as well as for different subpopulations of students.

a. *Capacity to discriminate between lowest and highest performers*

In order to evaluate whether one of these tests discriminates reading skills better than the other we compared the gap between the highest performers (upper third) and the lowest performers (bottom third) in each distribution of z-scores using the U-Mann Whitney independent rank-means comparison test. Results in both cases show significant differences between the bottom and upper thirds (Table 7), and suggest that both tests have similar discriminatory power.

Consistent with these results, we also find that most students stay in the same quartile of performance in both tests or move only one quartile up or down (around 81 percent, as shown in Table 8). There are only a few cases (4.4 percent) of students that perform very well in one test (first quartile) and very badly in the other test (fourth quartile), which denotes a difference in relative position of 3 quartiles from one test to another.

Table 7. Mean Comparison between lowest and highest performers in each test (Z-scores)

	Bottom Third			Upper Third			Z*	p-value
Test	**N**	**Mean**	**Std**	**N**	**Mean**	**Std**		
Oral test (Fluency)	158	-1.062352	.0443107	159	1.085425	.0451855	-15.415	0.0000
Written Test (Rasch Scores)	193	-1.138055	.0699292	147	.9544819	.0381903	-22.383	0.0000

* Z obtained using U-Mann Whitney test

Table 8. Difference in relative position (quartile) between performances in oral and written tests

	Difference (in quartiles)	Freq.	Percent	Cum.
Better position in written test (Rasch Scores)	3	13	2.74	29.9
	2	35	7.37	27.2
	1	94	19.79	19.8
Equal	**0**	**189**	**39.79**	**0**
Better position in oral test (Fluency)	1	99	20.84	20.8
	2	37	7.79	28.6
	3	8	1.68	30.3

It is possible that those few cases where students rank high in one test and low in the other share certain social or personal characteristics, indicating that one of the test may be behaving differentially for a given population subgroups.

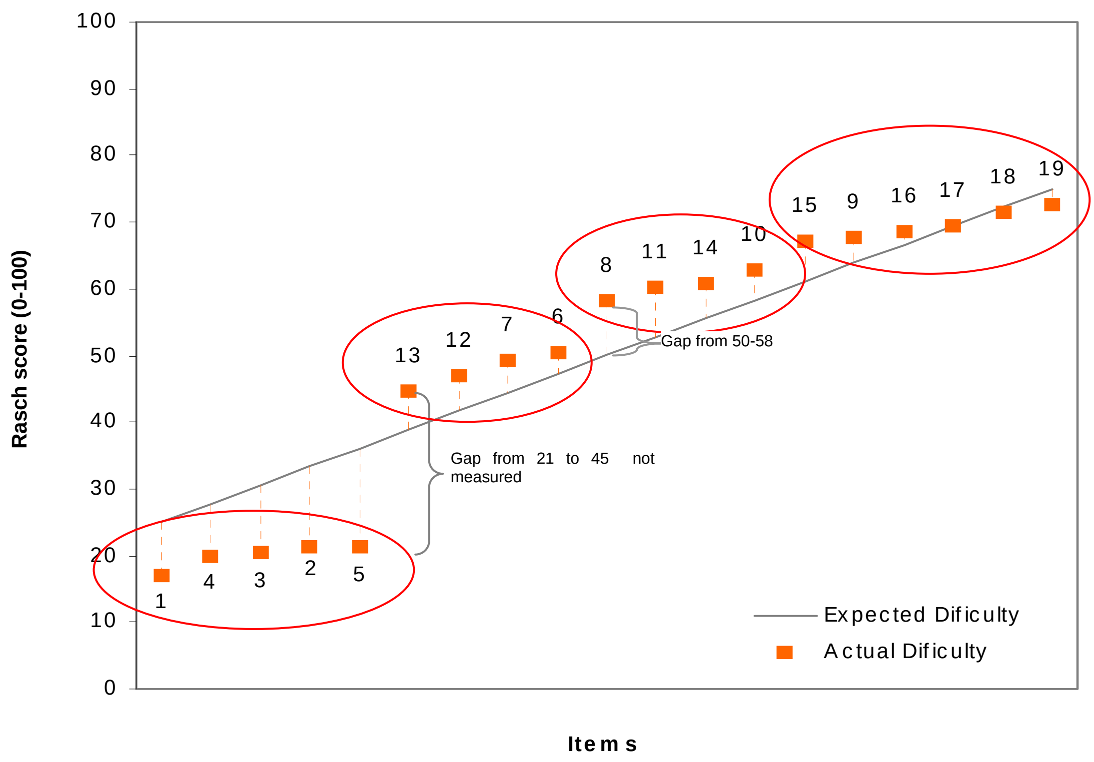
Figure 1 shows the estimated difficulty level of the ECE items in the sample using the Rasch model, compared to an ideal distribution of difficulty that shows increasing levels of difficulty by items going from 25 to 75 (mean ± 3 standard deviations). The figure reveals that there are roughly four levels of difficulty in the items of the test, closely associated with the type of questions. The first group of items (1-5) measures the ability of reading words and sentences. The second group comprises items measuring literal comprehension of short texts (a story and an informative poster). The third group includes items that measure inferential comprehension of the

[33] The Shapiro-Wilk statistic is the ratio of the best estimator of the variance to the usual corrected sum of squares estimator of the variance. The W values range from 0 to 1, and being close to one indicates normality of the variable. This test requires a sample greater than or equal to 7 and smaller than or equal to 2000.

short story and the poster. Finally, the fourth group measures the comprehension of a longer, descriptive text (and item 9, a complex inferential question on the short story).

These groups are similar to the description of the performance levels defined by MINEDU for this test, although the gaps between them reveal that the test may not be measuring performance levels between these defined groups. For instance, there is a large gap of unmeasured ability between the first and second groups of items. This indicates that the test, as it is, cannot discriminate among children with an ability level that falls between 21 and 44.7 in the scale. Incidentally, about 15 percent of the sample falls in this range. There is also a seemingly smaller gap between the second and third groups that does not allow discrimination among children scoring from 50 to 58, but over one third of the sample falls within this range. Furthermore, the order of items does not necessarily follow increasing levels of difficulty, and the first group of items turns out to be too easy for this population, as 97 percent of children are able to answer them correctly (see Annex 7.5).

Figure 1. Map of expected vs. actual difficulty of items using Rasch Scores for the ECE test (N=1001)



LEGEND
Words = items 1-3
Sentences = items 4 and 5
Short story= items 6-11
Poster (informative text) = items 12-14
Descriptive text = items 15 -19

Because of the measuring limitations mentioned above, this map gives us information for a heuristic interpretation, which is suggestive but not definitive. In order for the test to provide a good map of abilities in reading comprehension, items one through five should be more difficult so that not all children can answer them correctly, and not all should have the same level of difficulty. Also, there should be additional questions that discriminate ability levels between 21 and 45, and between 50 and 58. These limitations could be overcome using a larger number of items by rotated forms of testing (like the ones used by UMC in the sample evaluations of 2001 and 2004).[34]

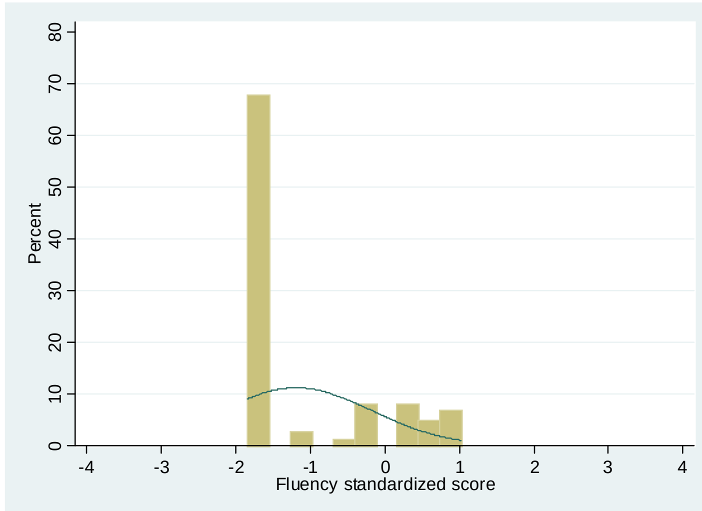
b. *Differential behavior of the tests across social groups: the case of Ashaninka students*

Even though there are no differences in the discrimination power between both tests, there is the possibility that they behave differently in different contexts or for different populations. This would mean that the tests show a differential behavior in some groups. When breaking down the results by characteristics of the students (gender, age first language, and access to pre-primary education) and schools (type and location) in the sample, we find no obvious differences between the two tests (see Annex 7.7), except in the case of Ashaninka students.[35]

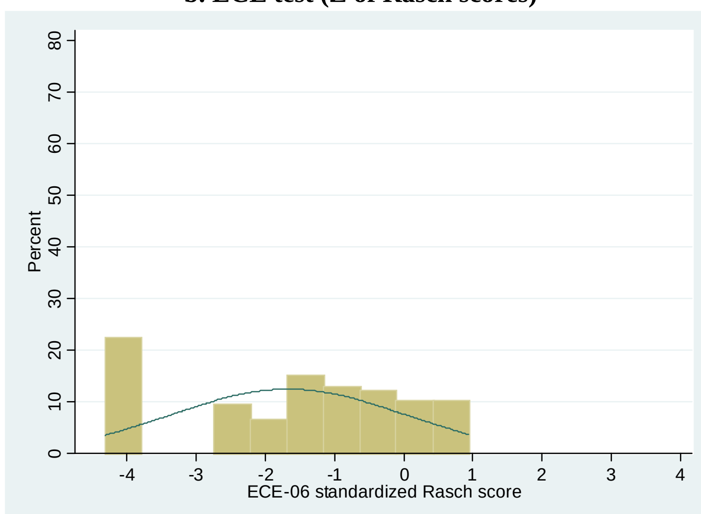
The fluency test does not seem to discriminate well among Ashaninka students: 68 percent of them score at the very bottom of the distribution, while the ECE test shows a more even distribution of scores, even among those performing badly (Figure 2).

Figure 2. Histograms of Z scores in both tests for Ashaninka students (N=40)

a. Oral Fluency test (Z scores)



b. ECE test (Z of Rasch scores)



The Kolmogorov-Smirnov test for two groups confirms this finding.[36] As shown in, the negative most extreme difference is larger in the oral test than in the written test. In other words, Ashaninka students are likely to under perform in the oral test.

[34] International tests such as PISA or TIMSS use hundreds of items distributed in several "rotated" forms of the same test, so that children taking Form A answer different questions than those taking Form B, but both answer a group of common questions. UMC has used this approach for the 2001 and 2004 evaluations. For instance, the 2004 evaluation used 153 items in second grade. See: (MINEDU-UMC, 2004).

[35] There were no differences between Spanish native speakers and non native speakers either.

[36] The two-sample KS test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples.

Three way Contingency Tables using Cochran-Mantel-Haenszel Method (CMH)[37] to test Differential Item Functioning (DIF) of the fluency dichotomic variable show that the odds of being classified as "fluent reader" in the oral test for Ashaninka students are only 5 percent relative to the odds of non Ashaninka students with the exact same reading ability (as measured in the ECE test). In other words, a student is 95 percent more likely to "fail" the fluency test if he or she is Ashaninka than if not, regardless of his/her reading ability. The differential effect of being Ashaninka on the fluency dichotomic measure is statistically significant (see Annex 7.8).

[37] The Cochran-Mantel-Haenszel (CMH) test compares two groups on a binary response, adjusting for control variables. The null hypothesis is that the response is conditionally independent of the "treatment" (group) in any given strata set according to the values of controlled variables. Mantel-Haenszel chi-square n. In statistics, an index of linear association between the row and column variables in a crosstabulation, calculated by multiplying the product-moment correlation coefficient by one less than the number of scores.

Table 9. Difference in score distributions between Ashaninka and Non Ashaninka students (Two-Sample Kolmogorov-Smirnov Test)

Statistics		Oral Test	Written test
N	Non Askaninka	435	437
	Ashaninka	40	40
	Total	475	477
Most Extreme Differences	Absolute	.647	.537
	Positive	.000	.000
	Negative	-.647	-.537
Kolmogorov-Smirnov Z		3.915	3.250
Asymp. Sig. (2-tailed)		.000	.000

4.2. What is the relationship between fluency and comprehension?

There are different ways to look at how fluency and comprehension measures relate. Most of the literature discusses bivariate correlation coefficients, and in some cases may present multivariate analysis to control for the effects of external factors. Here we do both, and also present a new way of looking at this relationship, using the Rasch model to transform fluency and comprehension scores into the same scale.

a. Correlation between fluency and comprehension measures

Figure 3 shows that there is a positive relationship between fluency and comprehension as measured by both tests. In this case, the Rasch scale helps illustrate the relationship more neatly, but regardless of which scale is used, it is clear that the more fluently a child can read, the higher she or he is likely to perform in the ECE written comprehension test.

Figure 3. Relationship between oral fluency and written comprehension (N=475, adjusted)

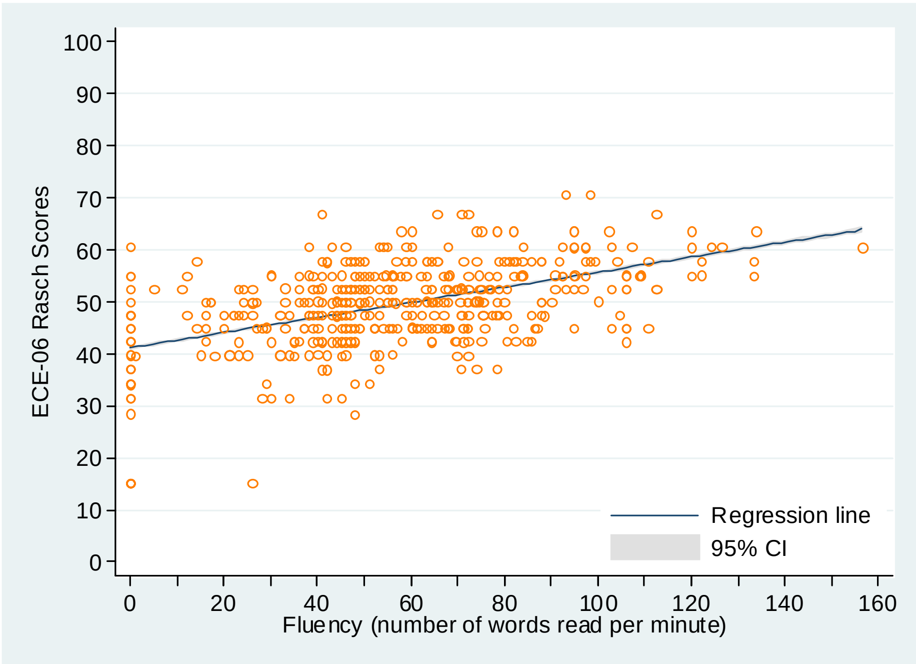


Table 10. Correlations between Fluency and Comprehension (Spearman Rho, N=475)

			Oral Test			Written Test
			Fluency	**Recall Comp.**	**Q&A Comp.**	**Rasch scores**
Individual Correlations	**Oral Test**	Fluency	1			
		Recall Comprehension**	0.3215*	1		
		Q&A Comprehension**	0.2592*	0.5724*	1	
	Written Test	Rasch scores	0.4705*	0.2386*	0.1999*	1
School Correlations ***	**Oral Test**	Fluency	1			
		Recall Comprehension**	0.3469*	1		
		Q&A Comprehension**	0.2160	0.3892*	1	
	Written Test	Rasch scores	0.5631*	0.3370*		1

* Sig. p<0.01. Coefficients that are not significant at p<0.1 are not shown in the table.
** N=431
*** Calculated from school means

To further evaluate the relationship between fluency and comprehension, we use a non-parametric correlation coefficient (Spearman Rho), because the distributions are not normal. We calculate the correlations both at the individual level and at the school level (Table 10). The relationship is significant and positive, as it was expected, but lower in magnitude than what has been found in other studies. There could be several explanations for this, including the sample size (much larger than other studies) and limitations of the written test in the sample. We will address this later in the discussion.

Another interesting –and perhaps more important– point is that the correlation between the fluency test and the written test total score (r=0.47) is higher than the average correlation between any item in the written test and the rest of the items in the written test (r=0.28), or than the average item-test correlation (r=0.4). Moreover, the correlation between oral fluency and the overall written test is higher than all but 3 out of the 19 item-test constructs in the written test (Table 32, in Annex 7.9). All this suggests that if a simple measure of oral fluency (or perhaps fluency using some other method) were to be considered part of a combined oral-written testing approach, the overall test would be perhaps superior to the written test *without* the oral test. In fact, when including a dichotomic fluency variable (45 words per minute) as an additional item in the ECE test the internal reliability increases slightly, from α=0.74 to α= 0.757 (see Annex 7.9), which indicates that the test has high internal consistency.

In addition, the correlation between comprehension in the oral test and comprehension in the written test is low, and fluency has a stronger correlation with comprehension in the written test than with comprehension in the oral test. This all suggests that this particular measure of oral comprehension may not have predictive nor concurrent validity, in the sense that it is not a good predictor of oral fluency and it does not relate strong enough with a different measure of the same construct.

In most cases, correlations at the school level are stronger than at the individual level, although the differences are not very high. While this difference could be explained in part by the fact that the school average eliminates outliers, it also points out the need for more analysis at the school level. On the other hand, the reduced and less significant correlations between the Q&A comprehension and the other variables can be explained by the also reduced variance of the Q&A scores (0.46 with a mean of 4.2, compared to 1.02 with a mean of 4.3 at the individual level).

Table 11. Correlations between Fluency and Comprehension for Ashaninka and Non Ashaninka students (Spearman Rho)

	Correlation coefficient
Ashaninka students (N=40)	0.6490*
Non-Ashaninka students (N=435)	0.3958*

* Sig. p<0.05. Coefficients that are not significant at p<0.1 are not shown in the table.

Now, given that the oral test behaves differently for Ashaninka and Non-Ashaninka students, it is also relevant to see how both tests correlate in these two groups. The Spearman's Rho coefficient for non-Ashaninka students is 0.396 between oral fluency and written comprehension, while for Ashaninka students is 0.649 (see Table 11).[38]

b. Relation between fluency and comprehension when social factors are taken into account

Despite the lack of normality in the distributions, we also run by multiple regressions to see how much of the variance in written comprehension is accounted for by oral fluency when controlling for relevant individual and socioeconomic variables. Table 12 shows that fluency in the oral tests explains over 28 percent of the variance in the ECE scores. The effect of fluency is significant even when keeping all other individual and school characteristics constant (age, Ashaninka language, preprimary education, the location and type of school), but the magnitude of the coefficient is slightly reduced.

Table 12. Regression Models (adjusted)

	Model 1	Model 2		Model 3	
Independent variables	**Coef.**	**Coef.**	**Beta**	**Coef.**	**Beta**
Fluency (oral test)	0.15 **	0.10 **	0.36	0.10 **	0.35
Full school		1.69 **	0.09		
Ashaninka		-7.51 **	-0.22	-13.28 **	-0.39
Rural		-1.22 **	-0.07		
Preschool education		0.35 **	0.02		
Age		-0.27 **	-0.03		
Fluency*Ashaninka				0.18 **	0.20
Constant	41.19 **	45.34 **	.	44.28 **	
Number of obs	475	473		473	
Prob > F	0.000	0.000		0.000	
R-squared	0.282	0.334		0.340	
Adj R-squared	0.282	0.334		0.339	
Root MSE	7.275	7.025		6.979	
VIF average	1.1	1.44		1.49	

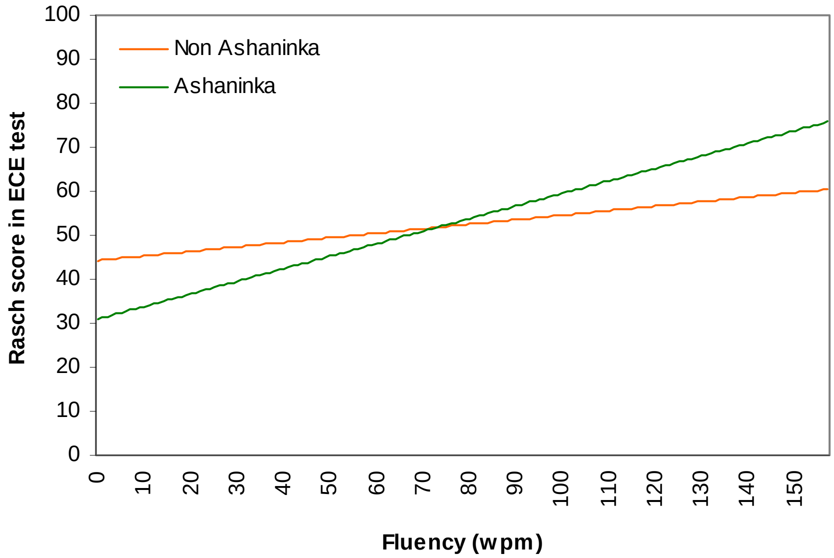
** p<.01
Coef. = Regression coefficient.
Beta = Standardized regression coefficient.

Moreover, the beta coefficient suggests that fluency contributes to explain much more of the variation in the written scores than any other single social or individual variable, except the Ashaninka language, which comes very close to fluency in terms of explaining variation in the written test. This suggests that the relationship between fluency and comprehension measures may be influenced by the language spoken by the child. Model 3 in Table 12 explores this matter

[38] There were not enough Ashaninka cases for obtaining correlation between fluency and oral comprehension measures.

further by incorporating a variable to account for the interaction between fluency and being Ashaninka. Figure 4 illustrates the results for Rasch scores, showing that the slope is steeper for Ashaninka students: in other words, for these students, the relationship between fluency and comprehension is stronger than for non-Ashaninka students (as also shown by correlation coefficients in Table 11). This may indicate that for the most disadvantaged children, working on fluency early on is even more important than with advantaged students, as it appears to have greater association with their ability to understand what they read.

Figure 4. Estimated relationship between Fluency and Comprehension measures for Ashaninka and Non Ashaninka students*

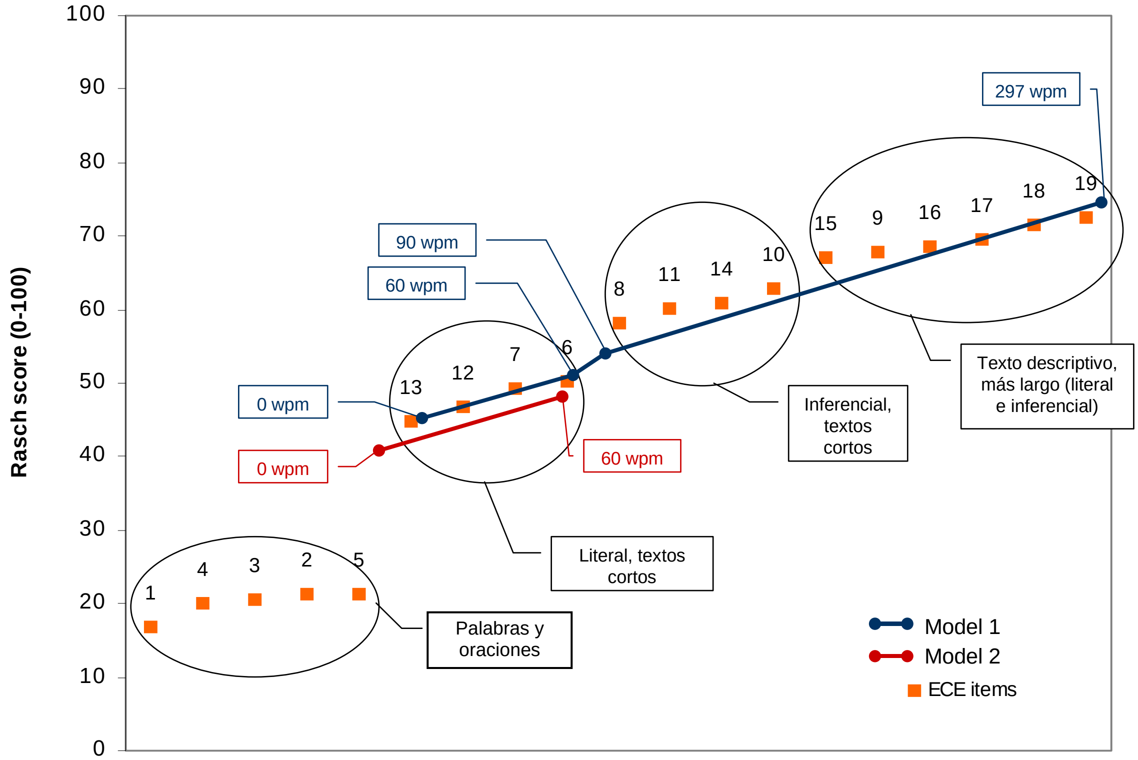


* Based on M6: Y= $\alpha + \beta_1$(Fluency) + β_2(Ashaninka) + β_3(Fluency*Ashaninka)

Having said that, an increase in the fluency test predicts only a small increase in performance in the written test:[39] Improving fluency by one word per minute is associated with only 0.15 in the Rasch scores (Model 1 in Table 12). That is, between a student who could not read a word aloud, and a student who read 60 words per minute, the difference in the ECE test would be only 9 points in the Rasch scale (from 41.2 to 50.2). When other factors are held constant (gender, age, first language, preschool education, location and type of school), this difference would be 6 points (from 45.3 to 51.3, in Model 2). Figure 5 illustrates this improvement against the estimated difficulty level of the questions. Holding other variables constant (Model 2), an improvement in fluency from 0 to 60 words per minute would mean an increase in likelihood of answering two or three more questions correctly, but the student would remain at the literal level. A student reading at 90 words per minute would still remain at that level. Basically, if using linear regressions to understand this relationship, we would have to say that in order to be likely to answer all questions correctly a student should be able to read at a fluency level of 297 words per minute. Of course, this is not very sensible: nowhere in the world is a second grader expected to read, aloud, at that speed, or a college graduate, for that matter. This suggests that the relationship between fluency and comprehension is likely non-linear and discontinuous.

[39] The possibility that this is due to the effect of multicolinearity is ruled out by the variance inflation factor (VIF), which is lower than 1.6 in all models.

Figure 5. Improvement in ECE scores with improvement in fluency (Models 1 and 2)



c. Is the fluency-comprehension relation non-linear or discontinuous?

The lower than expected results on both the correlations and regressions could indicate that the relationship between fluency and comprehension is not continuous but varies by level of performance. In other words, the fluency-comprehension relationship may be different for children who read more fluently than for those who don't, making it harder to estimate with linear predictions.

To explore signs of discontinuity in the fluency-comprehension relationship at or around the minimum rate of fluency needed for comprehension according to the literature, we compare the results in comprehension using two group variables: (1) a dichotomic variable that separates those who reach the minimum fluency level of 45 wpm and those who do not; and (2) an ordinal variable that groups students in quartiles of fluency.

Using boxplots, it is possible to see that the differences between fluency performance groups are more evident than correlation or linear regressions for the entire sample. Students who read 45 words per minute or more score significantly higher than those who read less fluently (Figure 6). The distribution of scores among the fluent readers shows also less variance and is less skewed than the distribution of scores of the non-fluent readers. When breaking down the sample in four performance groups, we can also see that there is a tendency to improve performance in the ECE test as fluency improves (Figure 7).

Figure 6. Box plots of Z scores in ECE tests in fluent and non-fluent readers (N=475, adjusted)

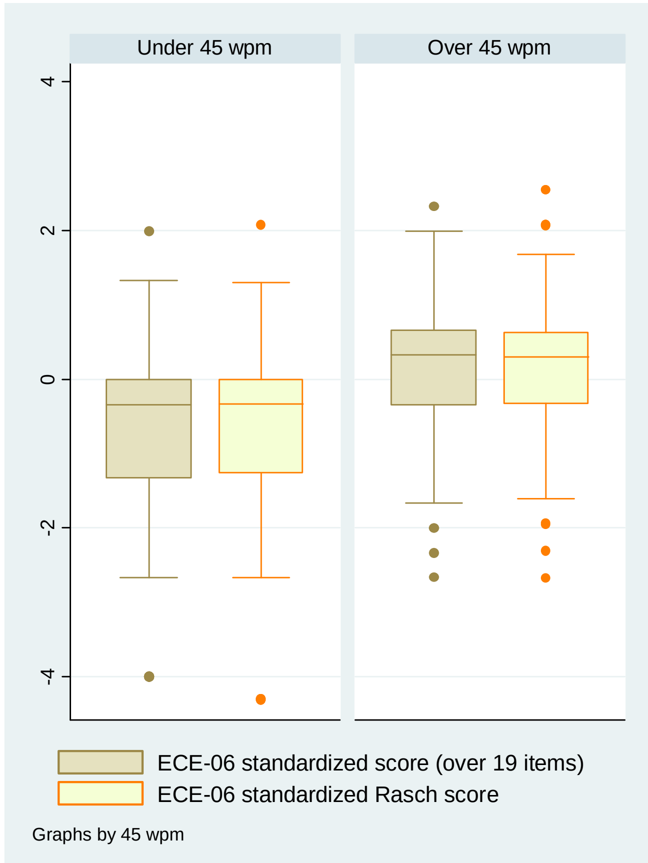
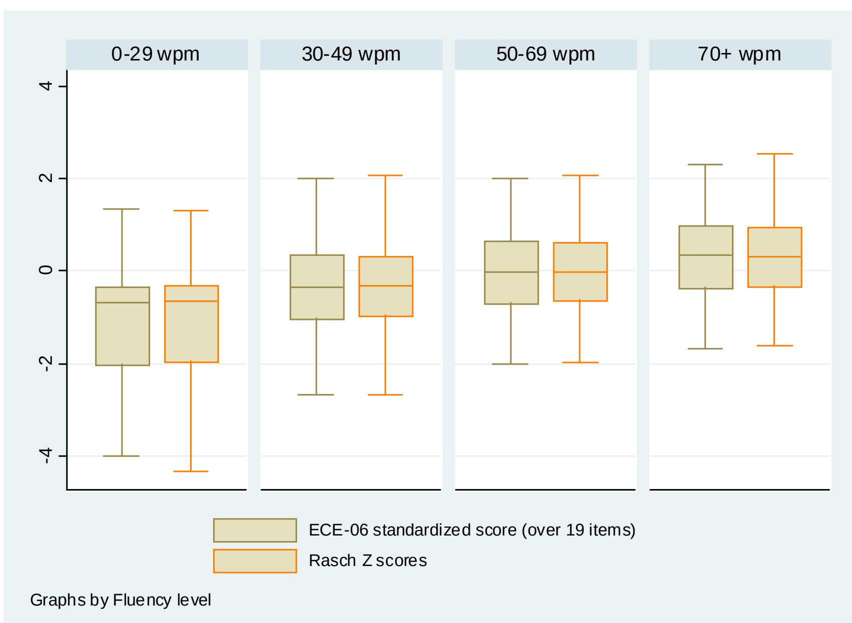


Figure 7. Box plots of Z scores in both tests by fluency quartiles (N=475, adjusted)



These results can be confirmed running additional non-parametric mean comparison tests to see if the differences between the fluency performance groups are statistically significant or not. For the dichotomic variable (45 words per minute) we use the U Mann Whitney test, a non-parametric equivalent to the t test that evaluates whether two independent samples are from the same population.[40] For the ordinal variable we use the Kruskal-Wallis one-way analysis of variance by ranks, a non-parametric method for testing equality of population medians among groups.[41]

Table 13. Mean comparison between minimally fluent and non-fluent readers with Rasch scores (U-Mann Whitney test)

	Rasch Scores		
	Obs	**Rank Sum (mean)**	**Expected Rank Sum**
Under 45 wpm	159	26,189 (165)	37,842
Over 45 wpm	316	86,861 (275)	75,208
Combined	475	113,050 (238)	113,050
Unadjusted variance			1,993,012
Adjustment for ties			-10.82
Adjusted variance			1,993,001.18
z			-8.254
Prob > |z|			0.000

The Mann-Whitney test shows that there are statistically significant differences in written comprehension scores between students who read fluently and those who do not (Table 13). What is more, such difference is not only significant but also large, as it is equivalent to nearly one standard deviation.[42]

Table 14. Differences in reading comprehension between fluency groups (Kruskal-Wallis test)

	Fluency Groups	N	Mean Rank
Ranks	0-29 wpm	85	134.22
	30-49 wpm	127	206.30
	50-69 wpm	117	261.62
	70+ wpm	146	307.07
	Total	**475**	
Test Statistics	Chi-Square		95.789
	df		3
	Asymp. Sig.		0.0001

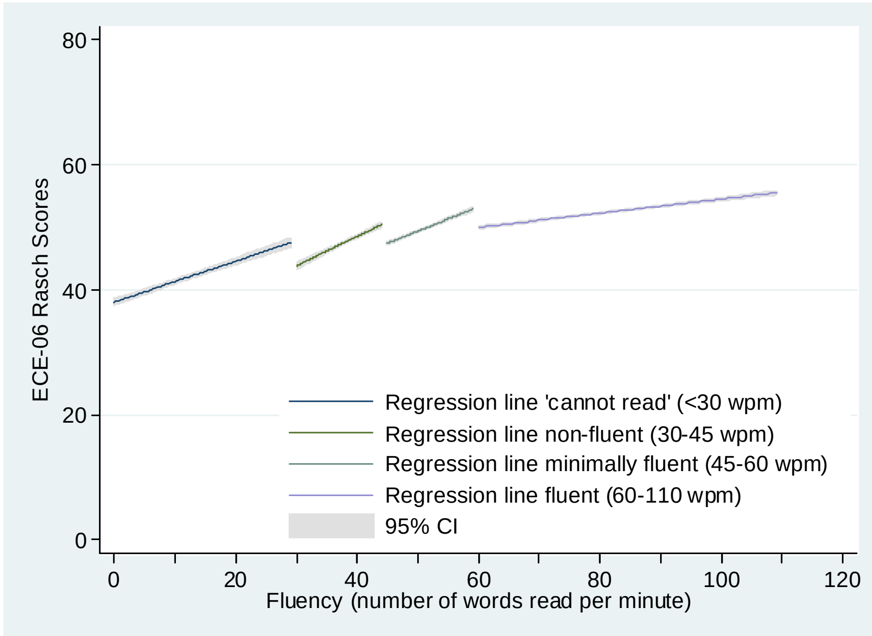
Breaking down the sample by quartiles of fluency we also find large and significant differences (Table 14): the mean rank in the bottom quartile is 134 while the mean rank in the upper quartile is 307. In other words, more fluent readers consistently rank higher in the written test than less fluent readers, and the probability that this would be due to chance is less than 0.0001 percent.

40 It is more powerful than the median test since it uses the ranks of the cases. Requires an ordinal level of measurement. U is the number of times a value in the first group precedes a value in the second group, when values are sorted in ascending order.

41 It is an extension of the Mann-Whitney U test to 3 or more groups.

42 Mean differences are also significant at other key cutpoints such as 60 and 75 words per minute. The differences using the 75 wpm cutpoint are higher than with the 60 wpm cutpoint, but both are lower than with the 45 wpm cutpoint (see Annex 7.10).

Figure 8. Effect of fluency on written comprehension by fluency level

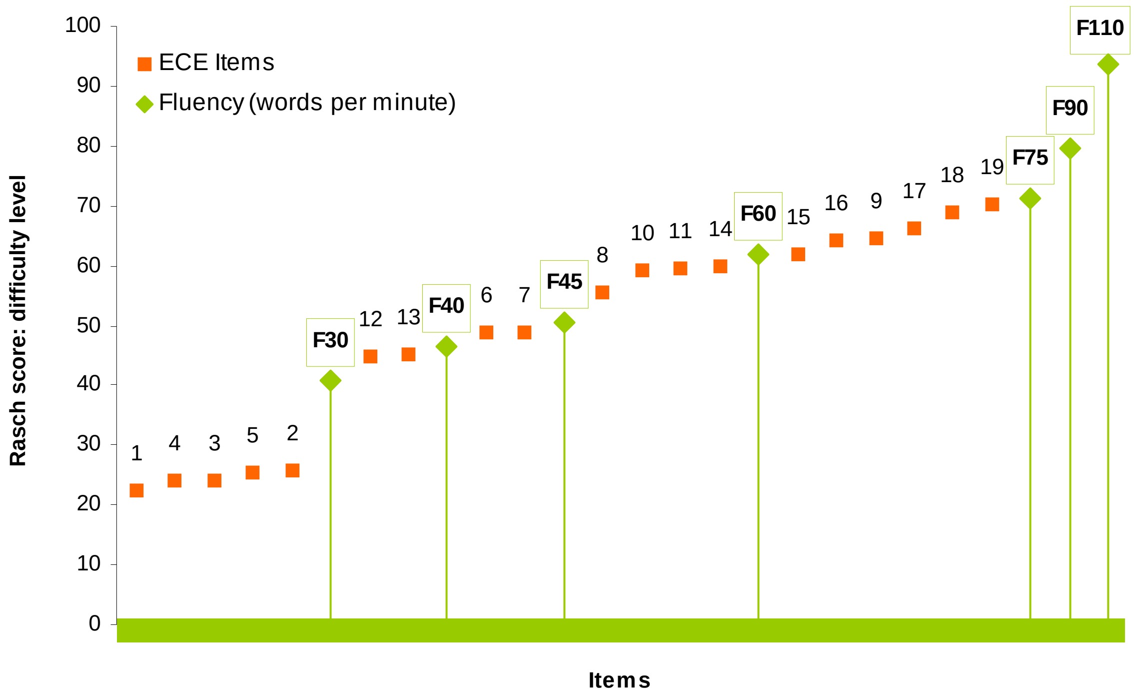


To illustrate this point, Figure 8 shows the effect of fluency on comprehension for different "fluency groups": when a student is starting to learn how to read (30 words per minute or less), the effect of fluency is strong. This effect gets stronger (i.e. steeper slope) when the student approaches the minimum level of fluency (30-45 wpm). From then on, the slope starts to get slightly less steep. After reaching 60 words per minute, the effect of fluency on comprehension is less pronounced.

d. A new way to look at the fluency-comprehension relation: application of Rasch scores

As noted before, one can also explore the relationship between the two tests by obtaining a map of difficulty level that includes all items of the ECE test and a measure of oral fluency in the Rasch scale. To do this analysis, we used several dichotomic versions of the fluency test, based on what has been identified in the literature as meaningful cut points of fluency: 30 words per minute, as the suggested standard for first grade in Peru (Crouch, Abadzi, Echegaray, Pasco, & Sampe, 2005); 40 words per minute, as the suggested cut point for fluency according to Davidson and Towner (unpublished); 45 words per minute, as the suggested cut point for fluency according to Abadzi (2006); and 60, 90 and 110 words per minute, as the suggested standards for second, third and fourth grades respectively in Peru (Crouch, Abadzi, Echegaray, Pasco, & Sampe, 2005). In addition, we estimated difficulty level at 75 words per minute, midpoint between 60 and 90 words per minute and also suggested benchmark for second graders in Spain (Equipo de Orientación Educativa de Marbella 2003; cited in Abadzi, 2006).

Figure 9. Map of difficulty of items using Rasch Scores and a dichotomic fluency variable* (N=475)



* The dichotomic variables distinguish between students who read at certain fluency levels. The number of words read per minute is noted in the label of the variable. For example, F30 is a dichotomic variable in which students who read 30 words per minute or more score 1, and those who read less score 0.

Seven dichotomic items are created to account for these various levels of fluency. We run seven separate analyses, one for each dichotomic item of fluency, to estimate the difficulty of different levels of fluency relative to the items in the ECE test. This was done with Sample B and the map is shown in Figure 9 (see Table 36 in Annex 7.4.b for detailed results).[43]

Results are quite suggestive. They indicate that a student that can read at least 30 words per minute (first grade level, as recommended for Peru), would be 100 percent likely to answer the first five items in the ECE test correctly:[44] that is, can read words and short isolated sentences. Then again, we've seen that any student, regardless of fluency, would be 96-97 percent likely to do so because these items are too easy.

A student who reads at least 40 words per minute can also understand a short informative poster at the literal level, with very high probability of answering items 12 and 13 correctly. This student, however, would be unlikely to answer correctly items 6 through 11, and 14 through 19. Here it is worth noting that the poster in question states prices of milk, cheese and butter and the days they are sold (see

43 The difficulty levels of the ECE items vary slightly from one analysis to the next (they get relatively easier as fluency level increases), but their order does not change. For illustration purposes, the difficulty levels of ECE items used in the plot come from their corresponding analyses: items 1-5, from the F30 analysis; items 12-13 from the F40 analysis; items 6-7, from the F45 analysis, items 8, 10, 11 and 14 from the F60 analysis; items 9 and 15-19 are an average of the difficulty levels estimated in the F75, F90, and F110.

44 Probability $= \frac{e^{(\theta-\beta)}}{1+e^{(\theta-\beta)}}$, where θ is the individual score, β is the item difficulty, and *e* is the exponential. In this example, θ=difficulty of F30= 41, and β= difficulty of item 2 = 25.9, then P = $\frac{e^{(41-25.9)}}{1+e^{(41-25.9)}}$ = 100.00.

Figure 10). The text is short and there are no full sentences or verbs, therefore requires less fluency than the short story (which has 112 words). Items 12 and 13 are literal questions regarding the information in the poster, while item 14 inquires on the purpose for which the poster was written, measuring inferential comprehension.

Figure 10. Informative Poster in ECE-06 (items 12-14 = questions 16-18)

Translation: "Every day FRESH MILK. Litter S/.2. Only Mondays and Fridays: Fresh cheese S/.8, Butter S/.5"

The cut point of 45 words per minute –which marks minimum fluency– falls exactly in the gap between the second and third groups of items in terms of difficulty (as were identified in Figure 1). These results indicate that minimally fluent readers are able to understand words and isolated sentences, and comprehend short texts at the literal level (highly likely to answer items 1-7, 12 and 13 correctly). Furthermore, if a student reaches the suggested fluency standard of 60 words per minute by the end of second grade, he or she should be able to at least understand words, sentences, short stories and informative posters at both literal and inferential level (i.e. answer correctly items 1-14, except 9), but this fluency rate does not yet guarantee that longer, descriptive texts are understood. In contrast, the benchmark suggested for second graders in Spain –a fluency rate of 75 words per minute– proves more difficult than all items in the ECE test, which suggests that a child reading at this rate or faster would be highly likely to ace the written test. Our analysis suggests that 75 words per minute would be the minimum fluency level required to reach full score in this particular ECE test, with at least 69 percent probability. Moreover, a student who is 100 percent likely to read at 75 words per minute (scoring 78.8 or higher) would be also 100 percent likely to answer all items correctly. Again, because of the limitations of the test to fit the Rasch model, this interpretation is more heuristic than statistical.

Finally, these results support the assumption that the relationship between fluency and comprehension measures is not linear: Contrary to the results of linear regressions (which indicate that with every 20 words increase in fluency the student is likely to answer one more item correctly), results based on Rasch analyses indicate that a five-word increase could lead to answering two more questions correctly. In fact, the impact of a five-word increase on comprehension is greater between 40 and 45 words per minute than between 30 and 35.[45] An increase by 15 words in fluency between 45 and 60 may lead to answering four more items correctly, while an equal increase between 60 and 75 would lead to answer six more items, marking the ability to understand longer, descriptive texts and answer inferential questions.

5. DISCUSSION

The findings of this paper open the way to rich discussions on several fronts: first, on the scopes and limitations of written and oral tests to measure reading skills; second, on the nature of the relationship between reading fluency and written comprehension; third, on the particularities of

[45] The difficulty of a dichotomic variable of 35 words per minute is estimated in 43.3, compared to 41 for F45, and no additional items in the ECE test are likely to be answered correctly.

measuring reading skills with disadvantaged students, taking as an example the Ashaninkas; and fourth, on how reading tests can be used to guide policymaking and strengthen accountability in the education system.

5.1. Scope and limitations of the two tests to measure reading ability

We found that the oral fluency test and the written comprehension test adequately measure students' reading abilities, and there are no differences in the discrimination power between both tests. However, there are some issues related to test construction and administration that need to be discussed.

With regards to the tests' construction, item and test analysis for both raw and Rasch scores of the written test revealed inadequate adjustment to psychometric assumptions in our sample: both scores have non-normal distributions, only moderate reliability, and lack unidimensionality; the first items are too easy for this sample, and the test needs more items of specific difficulty level to adequately measure all levels of reading abilities. In addition, the item analysis shows problems with several items, thought there are more problems using raw scores than with Rasch scores.

All of these problems could be explained in part by the limited number of items measuring second grade level abilities, as stated by curricular standards. The first items turned out to be too easy because they were measuring first-grade skills, while the other items that presented problems were actually measuring third and fourth grade skills.

However, the problems found pertain to this particular test (in this particular sample) and not to the written comprehension tests in general. The UMC's ECE tests can be a powerful tool to identify at-risk students and schools, as well as to track progress at the school level over the years, which would yield the necessary information needed to design and implement targeted programs and identify and reward successful educational experiences. The Rasch model makes it possible to establish equivalences with other standardized tests that are reported using the Rasch model, which, for instance, would allow tracking the progress of a child or a school over time. This is of great relevance now that the census evaluation will be done on an annual basis. All of this, provided that test developers in UMC design the test to fit the Rasch model from the start; include a greater number of items in rotated tests forms; and have enough time to pilot the items in a sample to make sure the test is unidimensional and reliable with an Alpha coefficient of at least 0.70, and provides an adequate mapping of skills based on the difficulty of the items.

The oral test could not be analyzed psychometrically, but there are no evidences of problems in the construction of the fluency measure. There are, however, indications of validity constraints with both comprehension measures (Q&A and Recall). Their low correlation with fluency and written comprehension indicates that they may have low predictive and concurrent validity. It seems that the Q&A comprehension measure is not able to adequately discriminate comprehension skills because most children were able to answer all the questions correctly. This led to little variance in the distribution of scores, which in turn decreased its correlations with other variables such as fluency and written comprehension. Indeed, Q&A Comprehension correlates the lowest with the rest, and these correlations become insignificant when done at the school level, as variance reduces even more. This is consistent with the results found in the broader Junin study (Banco Mundial, 2008), which led the authors to decide not to use Q&A Comprehension results. This problem could be solved by using a longer passage and adding more questions of greater difficulty. For instance, Fuchs, Fuchs, Hosp and Jenkins (2001) suggest ten questions per 400-word passage.

Recall comprehension, on the other hand, proved very difficult for the children. We found it to be highly disorienting for students who are used to repeating textually what the teacher or the material says. Most students in the sample tried to remember the passage word by word and many gave up after they could not go pass “Once upon a time.” Strictly, the real problem here is not so much that this comprehension measure assesses inappropriate skills –for it responds to curricular objectives– but rather the fact that in many public schools, teaching tends to encourage repetition rather than inferential or critical comprehension, as seen in classroom observations during the study (Banco Mundial, 2008). In any case, the recall comprehension measure tells us more about the gap between actual teaching practices and theoretical curricular guidelines, than any other single measure used in this study; yet, it is not the most appropriate for establishing the relationship between fluency and comprehension, as we will later discuss.

Regarding test administration, the team of evaluators was well trained and could control most administration conditions, that is: that children took the oral test in a private area with only the evaluators present, that they took the written test in silence and without interference of the teacher or any other school staff, and that they were given the time required to finish the tests. Test instructions were standard and written down for the evaluators to repeat. The evaluators personally made sure that classrooms and students were selected randomly and following the preestablished procedures. However, test administration could have been influenced by cultural, linguistic and racial differences between the evaluators and some of the students. In the future, it is suggested that at least one of the evaluators speaks the first language of the children being evaluated and is familiar with their culture, and that the tests are administered in both Spanish and the native language for the early grades.

Another problem in this study probably arises from the mismatch between the tests’ original target population –students at the end of second-grade– and the sample population –third-grade students in the second quarter of classes. This could explain, to some extent, why students could easily answer the first items of the written and the Q&A comprehension in the oral test. Fortunately, this is not an issue for future use of these tests, as the census evaluations are carried on at the end of the year and designed for each grade according to curricular standards. As for the oral test, it can be efficiently used to track progress among students in second and third grade during the school year by selecting passages with different lengths and complexity levels. If these oral measures are taken by the teacher or parents of the student, and if bad performance does not result in punishment of the child, then the contextual factors that influenced our sample could be overcome: the students would be familiar with their evaluator, could take the test in their native language, and eventually would grow familiar with this technique.

5.2. Nature of the relationship between oral fluency and written comprehension[46]

From the correlation results, it is possible to say that fluency explains 23 percent of the variance in written comprehension at the individual level and 32 percent at the school level in Junin, and that the probability of this result being due to chance is virtually nonexistent. This is considered a good effect size, and the effect of fluency on written comprehension is significant even when controlling for key individual and school characteristics, such as age, gender, first language, preschool education, type of school and location. In fact, it is by far the single most important variable in explaining performance on the written test. Furthermore, the fluency test tends to

[46] It is quite interesting that the correlation between fluency in the oral test and comprehension in the written test correlate higher than the oral comprehension measures correlate with oral fluency or written comprehension, but this has been discussed earlier, so here we will focus on the relationship between oral fluency and written comprehension.

explain more of the variation in the written test as well as in its items, than any item in the written test: the items in the written test explain on average only about 16 percent of the variation in the whole written test. The correlation coefficient between oral fluency and written comprehension is positive and highly significant (r=0.47, p<.0001), and there are several factors that could explain why it falls in the bottom end of the range found in other studies (between 0.40 and 0.91).

Why is the correlation between oral fluency and written comprehension not higher? There are several possible explanations for this. First, the effect of the sample size: the larger the sample, the smaller the coefficient gets, but also the more significant. Most of the correlation coefficients produced in the literature have come from small samples, between 15 and 70 cases (Fuchs, Fuchs, Hosp, & Jenkins, 2001; Shaw & Shaw, 2002), while our sample size was 475.

Second, given the moderate internal reliability of the ECE test, one would not expect to find stronger correlations with other measures of reading skills. As noted above, the internal correlations between the test items within the pencil-and-paper test were themselves actually worse than the correlation between the overall pencil-and-paper test and the oral test (average correlation between each written item and the rest of the items was only 0.34).

Third, the difficulty level of minimum fluency (45 wpm, Rasch score of 51) falls at around the mean difficulty level of the ECE test (Rasch score of 50) where there is a gap of "unmeasured ability." This ECE test does not have items measuring ability levels between scores of 50 to 58, but over one third of our sample falls in this gap. This means that precisely where the line that divides fluent and non-fluent children falls (and where most scores are concentrated), the ECE test is not able to adequately discriminate comprehension abilities.

Fourth, most of the international research relates fluency with literal comprehension, but it does not suggest that minimum fluency would be a good predictor (although it is certainly a requisite) for higher levels of comprehension, such as inferential comprehension, which was also evaluated in the ECE test. We find that the *minimum* fluency level of 45 words per minute has a lower difficulty level than half of the ECE test. In other words, minimum fluency would not be a good predictor of half the score in the ECE test (10 items out of 19). If we add the fact that the first five items are answered correctly by almost every student (hence, there is very little variability), there are only 4 items of the test that could be reasonably related to fluency. However, we find that a fluency level of 75 words per minute could be a good predictor of both literal and inferential comprehension of short and long second-grade level passages. Further research is needed to confirm this finding.

And finally, our results support the assumption that the relationship between fluency and comprehension measures is not linear. Hence, statistics such as correlations and linear regressions are not sufficient to obtain an accurate picture of the nature of this relationship. Instead, dividing the children into fluent and non-fluent using a cut-off point of 45 correct words per minute, very effectively discriminates between good and poor performers on the pencil-and-paper comprehension test. We find statistically significant differences between fluent and non-fluent readers in their performance on the written test. These differences are not only highly significant but also relatively large. The contrast between the moderate –but significant– correlation and regression coefficients, and the large –also significant– differences of means between fluent and non-fluent children indicate that the fluency-comprehension relationship is not continuous but might be, one could say, categorical (or at least non-linear). Moreover, alternate statistical methods based on the Rasch model indicate that the level of fluency affects the relationship between increasing fluency and increasing comprehension scores. For example, results suggest that the difference in comprehension between fluency rates of 40 to 45 words per minute may be greater than between rates of 30 to 35; it also may be greater between 60 and 75 than between rates of 45 and 60.

Why are the correlations stronger at the school level than at the individual level? Inter-school variation is larger in Peru (and in other countries with a high inequality) than in most developed countries. It is probably larger than intra-school variation. The broader Junin study (Banco Mundial, 2008) found that rural, multigrade schools that teach indigenous children are far behind than urban, full grade schools. There are many reasons for this: (a) rural schools have fewer resources such as appropriate text books and teaching materials; (b) teachers assigned to rural, multigrade schools are often less qualified and less experienced than those working in large, prestigious urban schools; (c) teachers in remote schools miss more hours of class than teachers in urban schools because they have to travel larger distances and no one controls their attendance; (d) teachers in multigrade classrooms must split their attention among groups of students, and students waste time waiting for the teacher to finish with another group and to come work with them; (e) children in urban schools have parents with higher level of education and more resources to spare than children in poor, rural communities; (f) in many rural communities children enter first grade without having had any exposure to written texts (no signs in the town, no books); and (g) indigenous children learn Spanish as they learn to read, and in Junin there are no teachers qualified on bilingual education.

Beyond the obvious differences between poor rural schools and urban schools, another likely source of school-to-school variation is instructional time use, which is a mediator variable that is challenging to measure, so it often escapes scrutiny. An international study found that students are often taught for only a fraction of the intended time, particularly in lower-income countries (Abadzi, 2007). In the fieldwork in Junin, classroom observations were carried on in the classes where we tested the students, estimating that in one hour of class, teachers were not engaged in teaching –were outside the room, working in their desks, talking to school personnel, or chatting with some students– an average of 10% of the time (16% in single-teacher schools). This is on top of the delays to start class, and teacher and student absenteeism (Banco Mundial, 2008). Research suggests that students must get sufficient time to process the information and practice to be able to learn the contents and develop the skills (Abadzi, 2007).

The fact that between-school variation is relatively large compared to within-school variation suggests that accountability for whole-school results, using school averages, might be both better (since accountability perhaps needs to address the worst forms of inequality first), and more reliable (since the correlations between items are higher across schools than across individuals). It is likely that whole-school accountability might have less perverse consequences than accountability over individual student results and be more effective in confronting inequality. In terms of policy implications, it is necessary to focus more on school variation than on individual variation.

A new framework for understanding the fluency-comprehension relation. Analyses based on the Rasch model to combine the ECE test and dichotomic versions of oral fluency (at several cutting points), have provided a more detailed description of the relationship between the individual items in the written test and different fluency levels. Results indicate that minimal fluency (45 wpm) is a good predictor of basic, literal comprehension, and a requisite for higher levels of comprehension but not a good predictor of performance at those levels. Yet, higher levels of fluency (e.g. 75 wpm) may be good predictors of more complex types of comprehension.

Foremost, this approach to examining the relationship between reading fluency and comprehension reveals that this relationship is likely to be non-linear and discontinuous. The strength of the relationship depends on the level of fluency and the difficulty of the comprehension questions. There may be critical points in the process of learning to read when stressing fluency in teaching may have a greater impact on improving of students' comprehension. In future and improved versions of these tests, Rasch analysis can help determine: (1) fluency rates that are likely to ensure good performance in a written test, (2) critical points in the early stages of reading where working on fluency would be likely to have a strong impact on comprehension, and (3) critical points later on where it would be more effective to shift the main focus of teaching reading to comprehension and vocabulary.

5.3. The Ashaninka case: a look at measuring reading with disadvantaged students

From all the social and individual variables analyzed, being from Ashaninka origin[47] has proven to be the single most important characteristic for explaining learning disadvantages (Banco Mundial, 2008). In fact, we find that it plays an important role in terms of performance in reading tests. First, Ashaninka students tend to underperform in the oral test, compared to their non-Ashaninka peers who scored similarly in the written test. And second, the relationship between fluency and comprehension is stronger for Ashaninka students than for non-Ashaninkas.

There are some test administration factors that can play a role in the performance of Ashaninka students in the tests. For instance, the fact that the evaluators had no previous contact with the children or their communities, were racially and culturally different from the students, did not speak their language, and the tests were administered in Spanish. All of these issues are likely to interfere with students whose first language is not Spanish and have limited contact with Spanish speaking people. So the same evaluator may generate more anxiety in Ashaninka children than in Spanish speakers because of the familiarity with the cultural codes and language. The effect of these variables seems to be stronger in the case of the oral test. Ashaninka students tend to under perform in the oral test –specifically, in fluency and recall comprehension– relative to the written test. This is likely to be the result of their limited fluency in Spanish and performance anxiety due to unfamiliar testing conditions. Reading aloud is more difficult than silent reading, generates more anxiety and self-awareness. Ashaninka children may be able to understand what they read but get blocked when asked to read it aloud because they do not feel comfortable with their pronunciation. Thus, some of those children who could not read one word aloud were able to answer correctly several questions in the written test.

Recall comprehension requires more oral language skills than direct Q&A (which demands one-word answers such as "yes," "no" or "Dogo"). As a result, the recall comprehension measure is more influenced by proficiency in Spanish than the Q&A measure, which creates validity problems and puts non-native Spanish speakers at a disadvantage. The fact that the fluency test does not seem to discriminate well among Ashaninka students, compared to the ECE test, suggests that the oral test can behave differently in certain population groups, which may reduce the external reliability.[48] The ECE test, on the other hand, appears to be less vulnerable to the influence of some contextual factors of the test administration itself. If oral fluency and recall comprehension were taken in both the first language of the students and Spanish by well-trained local evaluators, then we would have a more wholesome, reliable and valid measure of their

[47] The sample included students from four indigenous groups (Quechua, Aymara, Ashaninka, and Shipibo-Conibo) but only Quechua and Aymara had enough cases to run separate analysis.

[48] External reliability refers to the extent to which data measured at one time is consistent with data from the same variable measured at another time.

reading abilities. To prove this theory, a test-retest study would be needed under the aforementioned conditions.

From a psychosocial point of view, there is also the issue of stereotype threat as a factor influencing the performance of minority students. Our classroom observations revealed that teachers often send discouraging messages –in some cases this would be an understatement– to their students and this tends to happen more so when the teacher and students do not share the same cultural background. Reports of the use of psychological and physical threats and punishment by children and by our evaluators were more frequent in indigenous communities than in urban areas (Banco Mundial, 2008). Teachers in these schools also revealed in their interviews and in comments during class, low expectations to what their students could achieve academically. Thus, it is not far fetched to think that Ashaninka students will also have low expectations of changing the way their teachers or any foreign evaluator perceives them, or would feel very anxious when facing an evaluation in the fear of being punished. More research is needed to explore how stereotype threat influences performance of Peruvian indigenous children on standardized and non-standardized tests in Spanish.

Finally, the strength of the fluency-comprehension relationship can vary for groups of students that share certain characteristics, compared to the general population. For instance, international research has demonstrated that this relationship is stronger among students with reading disabilities (Hosp, 2001). Similarly, in our study, we found that the relationship is stronger for Ashaninka students: an increase in reading fluency translates into a larger increase in comprehension than it does for non-Ashaninka students, even when holding other factors constant. This may indicate that for the most disadvantaged children –the Ashaninkas being one example–, working on fluency early on is even more important than with advantaged students, as it appears to have greater impact on their ability to understand what they read. Also, this suggests that oral measures of reading abilities, such as fluency, could be appropriate tools used by teachers and parents for monitoring the progress of at-risk students in the early stages of learning to read, providing valuable information for lesson planning and for identifying individual students who are falling behind their peer group to make sure they catch up.

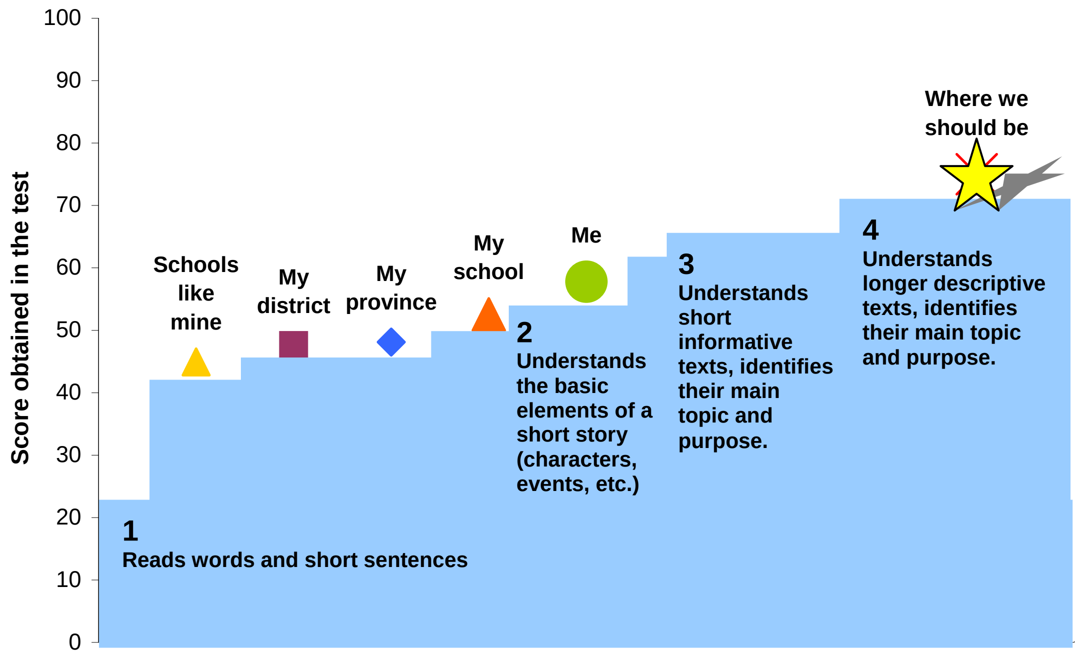
5.4. The multiple uses of reading tests: policy recommendations

This research was undertaken in the context of a broader body of work aimed at strengthening accountability through the use of results-based standards in education, health and nutrition. In education, the most important goal is for students to learn, and therefore, measuring student achievement is key to setting a meaningful standard and produce the information that will tell us how far are we from getting there. This paper was written to explore two specific measures of reading abilities, which are the foundation for any further learning.

In terms of reporting results, the Rasch model allows to present results both with references to norms (NRM) and to learning criteria (CRM). If done properly, it would give teachers, parents and administrators a more accurate idea of what precise skills were mastered by the students and how this performance compares to others. Raw scores or percentages of right answers do not provide that information.

Based on our data in the ECE test we propose a model for reporting results that can contribute to strengthening accountability, and then discuss how the reported results could be used specifically to define standards, mobilize the education community to reach these standards and keep them focused, track and communicate the progress along the way, identify those most in need of extra support, and strengthen accountability.

Figure 11. A model for reporting individual reading comprehension results in Rasch scores



A proposed model for reporting results in the ECE test. Figure 11 takes the case of one particular child to show his results in the ECE test, graphically comparing them to the averages obtained in his or her school, district, province and also in schools similar in type and location to his or her school,[49] while also showing how far he needs to go to reach the final "reading goal." These are placed in the same scale, illustrated as a "staircase" of abilities. The figure describes four critical achievements along the staircase: (1) ability to read words and short sentences; (2) ability to understand the basic elements of a short story (characters, events, etc.); (3) ability to understand short informative texts, identify their main topic and purpose; and the final goal for second grade, (4) ability to understand longer descriptive texts, identify their main topic and purpose.

The figure can be part of a report summarizing the results. The summary should have two separate pieces of information. First, a paragraph that tells the parent how the child is doing. And second, a paragraph that tells the parent how the school is doing compared to other similar schools. Something such as the following could be considered:

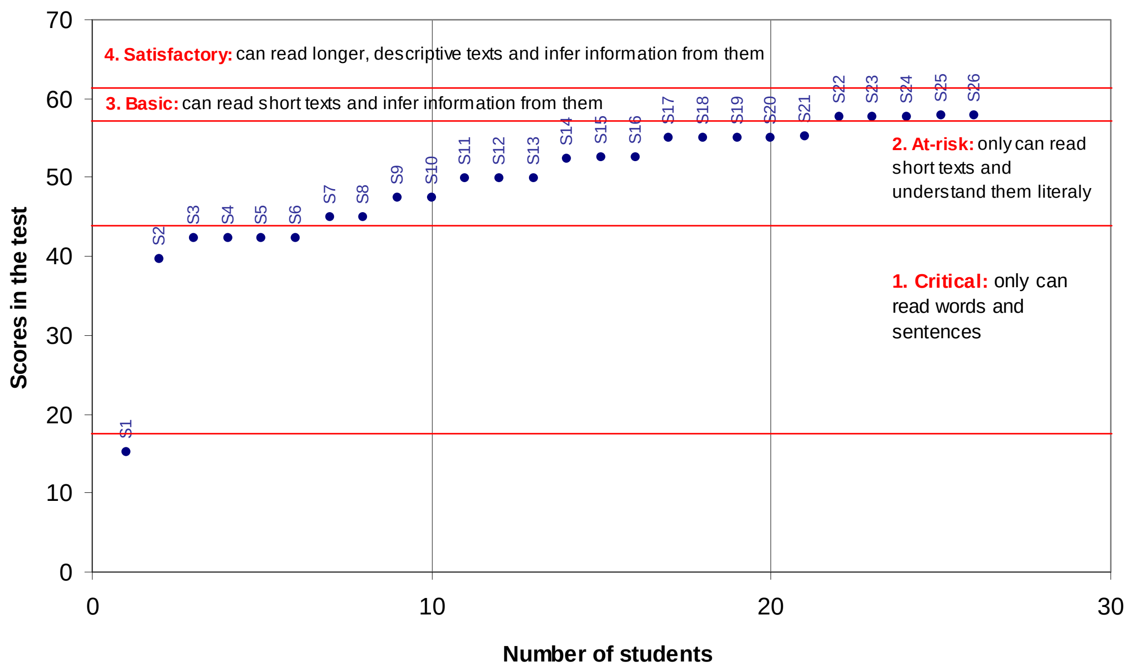
> *Dear parent,*
>
> Your child can read short stories or passages, and understand who are the main characters and what happens to them, but still needs to practice more to be able to read at the level expected for his grade. He should be able to read longer texts, and understand why they were written and what is their main message.

[49] "Schools like mine" distinguish five types of schools, combining type of school and location: Rural single-teacher, rural multigrade, rural full grade, urban multigrade and urban full grade. There are no urban single-teacher schools.

The school your child attends obtained better results than the average school in your district and province, and much better results than the average multigrade, rural school in the region. However, students in this school have not reached the goals set for their grade.

A similar graph could be used for teachers and principals, replacing "My class" instead of "Me," to compare the performance of one class relative to the entire grade. An additional plot could be added to show teachers where each individual student falls, suggesting that they focus on the students in critical situation or at-risk. Figure 12 shows an example. The important thing is to report results with clear markers that help easily identify different performance groups and where each student falls. Students are assigned a number (S1 through S32 in this case) and a list matching names and numbers is given to the teacher as a legend, along with simple instructions on how to work with students at each level of performance. Again, all of this is possible to do using Rasch scores, as long as the test fits the model. Raw scores, on the other hand, would not provide adequate information to produce this type of report.

Figure 12. Map of reading performance in one third grade class



Using reading tests to define standards. Clearly, the reality of Junin lags far behind from the curricular standards for reading comprehension in second grade. Using reports of individual and school-level results in the census evaluations in a format like the one suggested in this discussion could help illustrate the distance between the current situation and curricular goals. However, it could also turn out to be discouraging. Thus, it is important to set short-term (one-year) goals that could be reachable with the resources schools currently have, and then offer additional pedagogical support to those schools in the bottom quintile of performance to help them reach the goals. For a goal to have a mobilizing power, it needs to be: (1) clear and easy to understand by everyone involved (parents, students, teachers, administrators, authorities); (2) shared by all schools (school-specific goals can be set in addition to a general, common goal); and (3) easy enough so that all schools can potentially reach it, but hard enough to ensure progress will be made to close the gap between current performance and curricular standards. For instance, one form of standard could be that "every student in second grade would be able to read short stories and understand who are the characters and what happens to them." This is equivalent to reaching the second level showed in figures 10 and 11, still far from the curricular standard but ambitious enough (53 percent of students in our sample scored below level two).

Using reading tests to mobilize the community towards a common goal. Standards cannot be reached if there are no changes in strategies in the pedagogy and the use of instruction time: more hours of actual class work; culturally and linguistically appropriate materials and methods to teach reading; extracurricular activities such as "the reading hour," reading groups or pairs, etc. Some of these strategies can be implemented by the schools themselves, with participation and collaboration of the parents, others can be promoted at the local level, and others need to be part of regional or national policies. But in any case, even when the school community commits to implementing a series of activities to improve the students' reading skills the enthusiasm will fade away eventually if results are not seen fast enough.

Using reading tests to track progress and keep the focus. A standard that is evaluated only once a year will hardly mobilize the education community throughout the year. With any luck, it will stir things up right before the test and right after results are returned to the schools, but the "cool-off" period lasts for more than half of the school year. So it is important to also set progress goals that can be measured by the teachers and parents, and that provide quick feedback on the impact of specific measures that are being implemented. This will help them adjust strategies and also keep focused on the task.

This paper has shown that fluency is a good predictor of reading comprehension. Thus, if this were to be the goal to be reached at the end of the year –when the next census evaluation comes along– progress goals could be set in terms of fluency and recall comprehension: 45 words per minute and the child being able to identify the characters and events in the stoy. After reaching that goal, the standard can be raised to 60 words per minute and the child being able to tell what the story was about (the main topic and purpose of the text). And after that, it could be 90 words per minute and the child being able to give a critical opinion about the passage. And so on and so forth.

The measurement of fluency is relatively easy to do and somewhat mechanical, so anyone with minimum training and a watch can do it. At the early stages of learning to read, fluency is crucial and can only be developed by practicing. The test itself, more than a measure, is also an opportunity to practice. In addition, it gives the teachers information on which students require more attention and which ones are making good progress. It also gives parents information on the progress of their children at school. Even an illiterate parent is able to discern whether his or her child can or cannot read fluently, as the sound of a non-fluent reader is unmistakable, especially in comparison to the sound of a fluent reader. The typical sound of fluent readers can easily be disseminated by radio, for example, along with messages meant to encourage appropriate fluency and intonation, rather than speed at all cost. Once the child reaches a minimum level of fluency, an illiterate parent can also ask the child to explain the reading or to tell the story again, which is more in line with the oral story-telling tradition shared by most indigenous groups. Parents with low levels of education would not be able to produce closed questions in the spot but can simply ask "so, what was it about?" and "did you like it? Why?"

Measuring comprehension is trickier than measuring fluency, but even if there is no objective standard for it yet, oral reading tests like EGRA can contribute to improving comprehension and keeping the interest in reading. It is not advisable that the test be used to only coach for fluency. First, because it could have an adverse effect in the children, creating too much pressure to read fast and taking away the pleasure of reading. And second, the evidence that coaching in fluency is *causally* related to comprehension is not strong enough, either based on this study or the literature, to encourage fluency coaching as single-minded strategy in improving reading. However, in a context of developing skills in other areas (such as vocabulary, and a direct attack on comprehension), and when used as a marker or proxy, it can help detecting forward movement and focus attention on results. Children must be encouraged to discuss each reading, relate it to previous experiences or materials they are familiar with, give opinions to whether they liked it or not and explain why, and so on.

Using reading tests to communicate with stakeholders and strengthen accountability. In any case, it is crucial that results are communicated to parents, teachers, principals and the general public in a way that is easy to understand even for those with low levels of formal education. At the individual level, reports could inform parents how are their children doing (e.g. Figure 11) with a clear, short explanation. At the class level, reports should inform the teachers, how are their students doing (e.g. Figure 12) and clearly point out those who require extra help, laying out a set of strategies to apply with students at each level of performance. At the school level, reports should inform the entire community (parents, teachers, students, administrators, etc.) how the school is doing, including one figure similar to Figure 11 to show how it is performing compared to the averages in the district, province, region and country; and a table to track record of previous years to show how much progress the school made from one year to the next (Table 29). At the local and regional level, reports should inform authorities and education sector officials on how the schools in their jurisdiction are doing, including maybe a figure similar to Figure 12, but where the points are average scores per school instead of student scores. At the national level, reports should inform the MINEDU and MEF which jurisdictions are obtaining better results, which ones are falling behind, which are the most at-risk schools of the country and where are they located.

Table 15. Hypothetical comparative report card

	My school	Schools like mine	Mi district	My province	My region	PERU
2006	39.7	31.4	39.7	47.4	42.4	44.9
2007	39.7	34.4	39.8	49.9	47.4	47.4
2008	42.3	44.9	44.9	49.9	47.5	49.9
2009	45.0	45.0	45.0	50.0	47.6	50.0
2010	42.4	47.5	47.4	52.4	49.9	52.5
2011	39.7	47.5	47.4	50.0	49.9	55.1
2012	47.4	49.9	49.9	50.0	50.0	57.7
2013	47.5	45.1	57.8	50.0	50.0	60.5
2014	50.0	47.4	60.6	52.4	52.5	66.8
2015	50.0	47.5	60.6	60.6	52.5	70.6
2016	50.0	50.0	60.6	60.7	52.5	70.7
2017	50.0	47.4	60.6	60.6	52.5	70.7
2018	57.7	47.5	60.7	66.9	52.6	70.7
2019	57.7	49.9	63.5	70.6	55.1	70.7
2020	57.7	45.0	63.6	70.6	57.7	70.7
2021	**57.7**	**45.0**	**70.6**	**70.6**	**66.8**	**70.7**

Color	Level	Min.	Max.
	0. NADA	0.0	25.0
	1.CRITICO	25.0	40.0
	2. EN RIESGO	40.0	51.0
	3. BÁSICO	51.0	62.0
	4. SATISFACTORIO	62.0	76.0

All of these reports can then serve to strengthen horizontal and upward accountability, while downward accountability is strengthened by more information and control on the resources and support given to schools and jurisdictions most in need.

However, in reporting test results to stakeholders it is important to take into account that not all parents, or even teachers for that matter, would be able to understand tables and plots like the ones suggested above. Therefore, not all parents or teacher would be equally inclined to use test results for accountability purposes or for improving teaching processes and learning outcomes. Other communication strategies must be used to help parents with low levels of schooling take equal advantage of this information than those with more years of education. These could include radio and tv spots, and videos to be distributed across the schools, explaining in an easy way, and in the parents' native language, the purpose of these tests and how to interpret and use the results.

Using reading tests for further research. Finally, while implementing these policy recommendations, further research should be carried out to evaluate the impact of the policies and strategies being implemented, preferable with a semi-experimental, triangulating approach that pays attention to classroom practices, strategies implemented to improve reading, and additional support received by the school.

It is important to mention that the new versions of the ECE test have improved significantly compared to the one used for this paper. They include now twice as many items, which would allow for better psychometric properties of the test and for smaller measuring gaps in the scale. Thus, reading skills of second-grade students are likely to be more accurately assessed in the new versions of this test. In this context, the work presented in this paper could be used as a methodology to explore the relationship between reading comprehension and fluency.

It would also be interesting to study this relationship across several groups of students, such as: those attending private schools, other indigenous groups, immigrants in urban schools, those attending bilingual schools, and so on.

6. REFERENCES

Abadzi, H. (2006). *Efficient Learning for the Poor: Insights from the Frontier of Cognitive Neuroscience*. Washington, DC: World Bank.

Abadzi, H. (2007). *Absenteeism and Beyond: Instructional Time Loss and Consequences* (Policy Research Working Paper No. 4376). Washington, DC: The World Bank.

Armbruster, B. B., Lehr, F., & Osborn, J. (2003). *Put Reading First: The Research Building Blocks of Reading Instruction*. Washington, DC: Center for the Improvement of Early Reading Achievement (CIERA).

Baker, F., & Kim, S.-H. (2004). *Item Response Theory* (2nd ed.). New York: Marcel Dekker.

Banco Mundial. (2006). *Por una educación de calidad para el Perú: Estándares, rendición de cuentas y fortalecimiento de capacidades*. Washington D.C.: The World Bank.

Banco Mundial. (2008). *¿Qué puede hacer un Gobierno Regional para mejorar la educación? El Caso de Junín*. Lima: Banco Mundial (in Press).

Barksdale-Ladd, M. A. (2000). What's at stake in high-stakes testing: Teachers and parents speak out. *Journal of Teacher Education, 51*(5), 384-398.

Barrientos, A. M. R. (no date). Velocidad Lectora. Retrieved June, 2008, from www.rmm.cl/index_sub.php?id_contenido=1128&id_seccion=310&id_portal=75

Bazán, J., & Millones, J. (2002). Evaluación psicométrica de las pruebas CRECER 98. In J. Rodriguez & S. Vargas (Eds.), *Análisis de los Resultados y Metodología de las Pruebas Crecer 1998. Documento de trabajo 13* (pp. 171-195). Lima: MECEP-MINEDU.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. . In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Philadelphia: Lawrence Erlbaum Associates.

Camba, M. E. (2006). Los niveles de comprensión lectora [Electronic Version]. Retrieved 21 Junio 2007 from http://creativecommons.org/licenses/by-nc-sa/2.5/.

Carmines, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. London: Sage.

Christensen, K. B., & Bjorner, J. B. (2003). SAS macros for Rasch based latent variable modelling. *Tech. Rep, 03*(13).

CNE. (2006). *Proyecto Educativo Nacional al 2021: La educación que queremos para el Perú*. Lima: Consejo Nacional de Educación.

Congreso de la República, P. (2003). *Ley General de Educación N° 28044*. Lima: Cogreso de la Republica del Peru.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87-185.

Crouch, L. (2005). El Sector Educación: Estándares, Rendición de Cuentas y Apoyo. In D. Cotlear (Ed.), *Un Nuevo Contrato Social para el Perú: ¿Cómo lograr un país más educato, saludable y solidario?* . Washington D.C.: Banco Mundial.

Crouch, L. (2006). *Reading Fluency as an Operational Indicator in the Early Grades*. Washington, DC: USAID.

Crouch, L., Abadzi, H., Echegaray, M., Pasco, C., & Sampe, J. (2005). Monitoring Basic Skills Acquisition Through Rapid Learning Assessments: A Case Study from Peru. *Prospects, Vol. 35, No. 2 (2005, 35*(2).

Davidson, M., & Towner, J. *Evaluating An Oral Reading Fluency Cut-Score To Identify Second Grade Children Who Are At-Risk For Reading Failure*.Unpublished manuscript.

Deno, S. L., Mirkin, P. K., Chiang, B., & Lowry, L. (1980). *Relationships Among Simple Measures of Reading and Performance on Standardized Achievement Tests.* Minnesota: Institute for Research on Learning Disabilities, University of Minnesota.

Department of Education. (2002). *Revised National Curriculum Statement Grades R-9 (Schools)*. Pretoria: Department of Education.

Dorans, N. J., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23*(4), 355-368.

Emenogu, B. C., & Childs, R. A. (2005). Curriculum, Translation, and Differential Functioning of Measurement and Geometry Items. *Canadian Journal of Education, 28*(1 & 2), 128-146.

FairTest. (2007). *Criterion- and Standards- Referenced Tests*. Cambridge, MA: FairTest.

Fuchs, L. S., & Fuchs, D. (2002). *Progress Monitoring, Accountability, and LD Identification. Testimony to the President's Commission on Excellence in Special Education*. Nashville, TN: Vanderbilt University.

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.

Graves, M. F., Brunetti, G. J., & Slater, W. H. (1982). The reading vocabularies of primary-grade children of varying geographic and social backgrounds. . In J. A. H. L. A. Harris (Ed.), *New inquiries in reading research and instruction.* (pp. 99-104). Rochester,N.Y.: National Reading Conference.

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

Hess, K. K. (2006). *Reading Development & Assessment of Early Literacy: A Review of the Literature*. Salt Lake City: Utah Department of Education.

Hirsch Jr., E. D. (2003). Reading Comprehension Requires Knowledge of Words and the World: Scientific Insights into the Fourth-Grade Slump and the Nation's Stagnant Comprehension Scores. *AMERICAN EDUCATOR*(Spring), 10-44.

Holland, P. W., & Thayer, D. T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Hosp, M. K. (2001). *Variables that Affect the Correlation between Fluency and Accuracy with a Measure of Reading Comprehension*. Nashville, TN: Peabody College of Vanderbilt University.

Hudson, R. F., Lane, H. B., & Pullen, P. C. (2005). Reading fluency assessment and instruction: What, why, and how? *The Reading Teacher, 58*(8).

Hulme, C., Roodenrys, S., Brown, G., & Mercer, R. (1995). The role of long-term memory mechanisms in memory span. *British Journal of Psychology, 86*, 527-536.

International Reading Association. (1999). High-stakes assessments in reading: A position statement of the International Reading Association [Electronic Version] from http://www.reading.org/pdf/high_stakes.pdf.

Iowa Testing Programs. (2005). *Guidance for Developing District Policy and Rules on Test Use, Test Preparation, and Test Security for the Iowa Tests*. San Antonio, TX Riverside Publishing Company.

Jones, D. M. (2002). *The 7±2 Urban Legend.* Paper presented at the MISRA C Conference 2002.

Kame'enui, E. J., & Simmons, D. C. (1990). *Designing instructional strategies: The prevention of academic learning problems*. Columbus, OH: Merrill Publishing Company.

Kame'enui, E. J., & Simmons, D. C. (2000). *Planning and evaluation tool for effective schoolwide reading programs*. Eugene, OR: Institute for the Development of Educational Achievement. .

Kenya Institute of Education. (2002). *Primary Education Syllabus*. Nairobi: Kenya Institute of Education.

Kim, S.-H. (2001). An evaluation of a Markov chain Monte Carlo method for the Rasch model. Applied. *Applied Psychological Measurement, 25*(2), 163–176.

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

LLECE. (2008). *SERCE: Segundo Estudio Regional Comparativo y Explicativo. Los aprendizajes de los estudiantes de América Latina y el Caribe*. Santiago, Chile: OREALC/UNESCO.

Mapuranga, R., Dorans, N. J., & Middleton, K. (2008, March 23 to 28, 2008). *Review of Recent Developments in Differential Item Functioning*

Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), New York, NY.

McKenna, M., & Stahl, S. (2003). *Assessment for reading instruction*. New York, NY: Guilford Press.

MED. (2005). *Diseño Curricular Nacional de Educación Básica Regular: Proceso de Articulación*. Lima: Ministerio de Educación, DINEIP-DINESST.

MEF. (2007). Transparencia Económica - Perú: Consulta Amigable de la Ejecución del Gasto Público (SIAF-SP) Retrieved January 2007, from http://transparencia-economica.mef.gob.pe/amigable/

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81-97.

MINEDU-UMC. (2004). Marco de Trabajo de las pruebas de Rendimiento. In *Evaluacion Nacional 2004.* Lima: Ministry of Education.

MINEDU-UMC. (2007). *Informe: Evaluación Census de Estudiantes 2006*. Lima: Ministerio de Educación. Unidad de Medición de la Calidad Educativa.

MINEDU. (2005). *Diseño Curricular Nacional de Educación Básica Regular: Proceso de Articulación*. Lima: Ministerio de Educación, DINEIP-DINESST.

Nagy, W. E., & Scott, J. (2000). Vocabulary Processes. In M. e. a. Kamil (Ed.), *Handbook of Reading Research* (Vol. III). Mahwah, N.J: Erlbaum.

Nelson, J. R., & Stage, S. A. (2007). Fostering the Development of Vocabulary Knowledge and Reading Comprehension Though Contextually-Based Multiple Meaning Vocabulary Instruction. *Education and Treatment of Children, Vol. 30, N° 1*, 1-22.

NRP. (2000). *Teaching Children to Read: An Evidence-Based Assessment of the Scientific Research Literature on Reading and Its Implications for Reading Instruction*. Washington, DC: National Reading Panel

OECD. (2007). *PISA – The OECD Programme For International Student Assessment*. Paris: OECD.

Orlando, M., Sherbourne, C. D., & Thissen, D. (2000). Summed-score linking using item response theory: Application to depression measurement. *Psychological Assessment, 12*(3), 354-359.

Paxson, C., & Schady, N. (2005). *Cognitive development among young children in Ecuador: The roles of wealth, health and parenting* (Policy Research Working Paper No. 3605). Washington DC: World Bank.

PNUD. (2006). *Índice de Desarrollo Humano Distrital 2005 (base de datos)*. Lima: PNUD.

RTI International. (2008). *EGRA: Frequently Asked Questions* (EdData II Technical and Managerial Assistance No. 3). Washington, DC: USAID.

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

Shaw, R., & Shaw, D. (2002). *DIBELS Oral Reading Fluency-Based Indicators of Third Grade Reading Skills for Colorado State Assessment Program (CSAP). (Technical Report)*. Eugene, OR.

Shealy, R., & Stout, W. F. (1993). A model-based standardization approach that separates true bias/DIF from group differences and detects test bias/DTF as well as item bias/DIF. . *Psychometrika, 58*, 159-194.

Shinn, M. R., & Shinn, M. M. (2002). *AIMSweb Training Workbook*. Eden Prairie, MN: Edformation.

Skiba, R. J., Simmons, A. B., Ritter, S., Kohler, K. R., & Wu, T. C. (2003). The psychology of disproportionality: Minority placement in context. *Multiple Voices for Ethnically Diverse Exceptional Learners, 6*, 27-40.

Smith, M. K. (1941). Measurement of the size of general English vocabulary through the elementary grades and high school. *Genetic Psychological Monographs*, *24*, 311-345.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African-Americans. *Journal of Personality and Social Psychology*, *69*(5), 797-811.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Taylor, B. M., Peterson, D., Rodriguez, M. C., & Pearson, P. D. (2002). *The CIERA School Change Project: Supporting Schools as They Implement Home-Grown Reading Reform* (CIERA Report No. 2-016). Ann Arbor. MI: Center for the Improvement of Early Reading Achievement, University of Michigan School of Education.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability *Journalism Quarterly*, *30*, 415-433.

The Access Center. (2005). Early Reading Assessment: A Guiding Tool for Instruction. Retrieved June, 2008, from http://www.ldonline.org/article/14510?theme=print

The Psychological Corporation. (2003). *Early Reading Diagnostic Assessment - 2nd Edition (ERDA-2)*. San Antonio, TX: Harcourt Assessment, Inc.

Thorne, C. (2005). Contexto sociocultural, desarrollo del niño y lectura inicial en el Perú. *Revista de Psicología de la PUCP, 23*, 139-163.

Torgesen, J., Wagner, R., & Rashotte, C. (1999). *Test of Word Reading Efficiency (TOWRE)*. Austin, TX ProEd.

University of Oregon. (2005). *Dynamic Indicators of Basic Early Literacy Skills -- 6th Edition (DIBELS-6)*. Oregon: University of Oregon.

Vaughn, S., & Linan-Thompson, S. (2004). *Research-Based Methods of Reading Instruction Grades K-3*: Association for Supervision and Curriculum Development.

Wiederholt, J. L., & Bryant, B. (2001). *Gray Oral Reading Test-4th Edition (GORT-4)*. Austin, TX: ProEd.

Williams, K. T. (2001). *Group Reading Assessment and Diagnostic Evaluation (GRADE)*. Circle Pines, MN American Guidance Service.

Wilson, M. (2004). *Constructing Measures: An Item Response Modeling Approach*. Philadelphia: Lawrence Erlbaum Associates.

WINSTEPS. Winsteps Help for Rasch Analysis: Reliability and separation of measures [Electronic Version]. Retrieved May 13, 2008 from http://www.winsteps.com/winman/index.htm?reliability.htm.

Wood, F. B., Hill, D. F., Meyer, M. S., & Flowers, D. L. (2005). Predictive Assessment of Reading. *Annals of Dyslexia, 55 n°2 2005*, 193-216.

Wren, S. (2004). Descriptions of early reading assessments. Retrieved June, 2008, from http://www.balancedreading.com/assessment/assessment.pdf

Wright, B. D., & Stone, M. H. (1979). *Best Test Design. Rasch Measurement*. Chicago: MESA Press.

7. ANNEXES

7.1. List of Available Reading Assessments

Source: SEDL (http://www.sedl.org/reading/rad/chart.html)

		Cognitive Elements Evaluated by this Assessment													
Criterion Based Assessments	Grades	Reading Comprehension	Language Comprehension	Background Knowledge	Linguistic Knowledge	Phonology	Syntax	Semantics	Decoding	Cipher Knowledge	Lexical Knowledge	Phoneme awareness	Alphabetic Principle	Letter Knowledge	Concepts about print
Abecedarian Reading Assessment	K and 1							✓	✓	✓	✓	✓	✓	✓	
Analytical Reading Inventory (7th Edition)	K, 1, 2, 3, and higher	✓							✓						
Assessment of Literacy and Language (ALL)	Pre-K, K, and 1		✓				✓	✓		✓				✓	✓
Auditory Analysis Tests	1, 2, 3, and higher											✓			
Bader Reading and Language Inventory - 5th Edition	Pre-K, K, 1, 2, 3, and higher	✓	✓			✓	✓		✓	✓		✓		✓	✓
Basic Reading Inventory -- 9th Edition	Pre-K, K, 1, 2, 3 and higher					✓			✓			✓		✓	✓
Bracken Basic Concept Scale - Revised	Pre-K, K, 1, 2							✓						✓	
Bracken School Readiness Assessment (BSRA)	Pre-K, K, 1, 2							✓							
Brigance Diagnostic Comprehensive Inventory of Basic Skills - Revised (CIBS-R)	Pre-K, K, 1, 2, 3, and higher	✓	✓			✓		✓	✓	✓				✓	
Brigance Diagnostic Inventory of Early Development II (IED-II)	Pre-K, K, 1, 2			✓				✓	✓	✓				✓	
Brigance K & 1 Screen II	Pre-K		✓				✓	✓	✓	✓				✓	
Brigance Preschool Screen II	Pre-K		✓				✓	✓							
Criterion Test of Basic Skills (CTOBS-2)	1, 2, 3, and higher								✓	✓					
Decoding Skills Test	1, 2, 3, and higher								✓	✓					
Developmental Reading Assessment, K-3 - 2nd Edition (DRA-2, K-3)	K, 1, 2, 3	✓						✓	✓	✓		✓		✓	
Diagnostic Assessments of Reading -- 2nd Edition	K, 1, 2, 3, and higher	✓						✓	✓	✓		✓		✓	✓
Domine Reading and Writing Assessment Portfolio (Revised)	K, 1, 2, 3, and higher								✓	✓		✓		✓	✓
Dynamic Indicators of Basic Early Literacy Skills -- 6th Edition (DIBELS-6)	Pre-K, K, 1, 2, 3	✓							✓	✓		✓		✓	
Early Reading Diagnostic Assessment - 2nd Edition (ERDA-2)	K, 1, 2, 3	✓	✓					✓	✓	✓				✓	✓
Ekwall / Shanker Reading Inventory -- 4th Edition (ESRI-4)	1, 2, 3, and higher	✓	✓						✓	✓		✓		✓	✓
Fluency: Strategies & Assessments (2nd Edition)	1, 2, 3, 4, 5, 6, 7, 8								✓						
Gray Silent Reading Tests (GSRT)	2, 3, and higher	✓													
Kindergarten Readiness Test (KRT)	K and 1 (Early 1st grade		✓					✓						✓	

Criterion Based Assessments	Grades	Cognitive Elements Evaluated by this Assessment													
		Reading Comprehension	**Language Comprehension**	**Background Knowledge**	**Linguistic Knowledge**	**Phonology**	**Syntax**	**Semantics**	**Decoding**	**Cipher Knowledge**	**Lexical Knowledge**	**Phoneme awareness**	**Alphabetic Principle**	**Letter Knowledge**	**Concepts about print**
	only)														
Observation Survey of Early Literacy Achievement, Revised 2nd Edition	K, 1, 2, and 3								✓					✓	✓
Phonemic Awareness in Young Children: A Classroom Curriculum	K and 1											✓			
Phonemic-Awareness Skills Screening (PASS)	1 and 2											✓			
Phonological Awareness Literacy Screening - 1-3 (PALS-1-3)	1, 2, 3	✓							✓	✓		✓		✓	
Phonological Awareness Literacy Screening - K (PALS-K)	K								✓	✓				✓	✓
Phonological Awareness Literacy Screening PreK (PALS-PreK)	Pre-K and early K									✓				✓	✓
Pre-Literacy Skills Screening (PLSS)	K and 1													✓	
Process Assessment of the Learner (PAL) -- Test Battery for Reading and Writing	K, 1, 2, 3, and higher		✓						✓	✓		✓		✓	
Rapid Automatized Naming and Rapid Alternating Stimulus Tests (RAN/RAS)	K, 1, 2, 3, and higher													✓	
Reading Analysis and Prescription System (RAPS)	K, 1, 2, 3, and higher									✓					
Reading Inventory for the Classroom (RIC) & Tutorial Audiotape Package- 5th Edition	Pre-K, 1, 2, 3, and higher	✓	✓						✓						
Ready to Learn: A Dyslexia Screener	Pre-K, K, 1					✓		✓						✓	
Rigby Reads (Reading Evaluation and Diagnostic System)	K, 1, 2, 3, and higher	✓					✓	✓		✓				✓	
School Readiness Test (SRT)	K and 1 (Early 1st grade only)		✓					✓						✓	
Stanford Diagnostic Reading Test -- 4th Edition (SDRT-4)	K, 1, 2, 3, and higher		✓					✓							
Stanford English / Spanish Language Proficiency Test (Stanford ELP / SLP)	K, 1, 2, 3, and higher	✓	✓												
Stanford Reading First	K, 1, 2, 3, and higher	✓	✓					✓	✓						
STAR Early Literacy Computer-Adaptive Diagnostic Assessment	Pre-K, K, 1, 2, and 3	✓		✓				✓	✓	✓			✓	✓	✓
STAR Reading Computer-Adaptive Reading Test	1, 2, 3 and higher	✓													
Stieglitz Informal Reading Inventory -- 3rd Edition	1, 2, 3, and higher	✓	✓						✓			✓			
Teaching Beginning Readers: Linking Assessment and Instruction	K, 1, 2					✓			✓	✓		✓		✓	
Tejas Lee	K, 1, 2, and 3	✓	✓						✓	✓				✓	✓
Texas Primary Reading Inventory	K, 1, 2, and 3	✓	✓						✓			✓			✓
The Critical Reading Inventory	Pre-K, K, 1, 2, 3, and higher	✓							✓						
Yopp-Singer Test of Phoneme Segmentation	Pre-K, K, and 1											✓			

Norm Reference Assessments	Grades	Cognitive Elements Evaluated by this Assessment: Reading Comprehension	Language Comprehension	Background Knowledge	Linguistic Knowledge	Phonology	Syntax	Semantics	Decoding	Cipher Knowledge	Lexical Knowledge	Phoneme awareness	Alphabetic Principle	Letter Knowledge	Concepts about print
Aprenda: La Preuba de Logros en Espanol - 3rd Edition	K, 1, 2, 3, and higher	✓	✓					✓							
Assessment of Literacy and Language (ALL)	Pre-K, K, and 1		✓				✓	✓		✓				✓	✓
Boehm Test of Basic Concepts - 3rd Edition (Boehm-3)	K, 1, and 2							✓							
Boehm Test of Basic Concepts - 3rd Edition - Preschool (Boehm-3)	Pre-K and K							✓							
Bracken Basic Concept Scale - Revised	Pre-K, K, 1, 2							✓						✓	
Bracken School Readiness Assessment (BSRA)	Pre-K, K, 1, 2							✓							
Brigance Diagnostic Comprehensive Inventory of Basic Skills - Revised (CIBS-R)	Pre-K, K, 1, 2, 3, and higher	✓	✓			✓		✓	✓	✓				✓	
Brigance Diagnostic Inventory of Early Development II (IED-II)	Pre-K, K, 1, 2			✓				✓	✓	✓				✓	
Brigance K & 1 Screen II	Pre-K		✓				✓	✓	✓	✓				✓	
Brigance Preschool Screen II	Pre-K		✓				✓	✓							
Comprehensive Test of Phonological Processing (CTOPP)	K, 1, 2, 3 and higher											✓			
Comprehensive Test of Phonological Processing (CTOPP)	K, 1, 2, 3											✓			
Diagnostic Assessments of Reading -- 2nd Edition	K, 1, 2, 3, and higher	✓						✓	✓	✓		✓		✓	✓
Domine Reading and Writing Assessment Portfolio (Revised)	K, 1, 2, 3, and higher								✓	✓		✓		✓	✓
Early Reading Diagnostic Assessment - 2nd Edition (ERDA-2)	K, 1, 2, 3	✓	✓					✓	✓	✓				✓	✓
Expressive One-Word Picture Vocabulary Test (EOWPVT)	Pre-K, K, 1, 2, 3, and higher							✓							
Gates-MacGinitie Reading Tests, 4th Edition (GMRT-4)	Pre-K, k, and 1, 2, 3, and higher	✓	✓							✓				✓	✓
Gray Diagnostic Reading Tests - 2nd Edition (GDRT-2)	K, 1, 2, 3, and higher	✓						✓	✓					✓	✓
Gray Oral Reading Test-4th Edition (GORT-4)	2, 3, and higher	✓							✓						
Gray Silent Reading Tests (GSRT)	2, 3, and higher	✓													
Group Reading Assessment and Diagnostic Evaluation (GRADE)	Pre-K, K, 1, 2, 3, and higher	✓	✓					✓	✓	✓				✓	
Kindergarten Readiness Test (KRT)	K and 1 (Early 1st grade only)		✓					✓						✓	
Lindamood Auditory Conceptualization Test - 3rd Edition (LAC-3)	K, 1, 2, 3, and higher											✓			
Metropolitan Achievement Tests, Reading Diagnostic Tests - 8th Edition (MAT-8)	K, 1, 2, 3, and higher	✓	✓					✓		✓					
Oral and Written Language Scales (OWLS)	K, 1, 2, 3, and higher		✓												
Phonics-Based Reading Test (PRT)	1, 2, 3, and higher								✓	✓					
Pre-Reading Inventory of Phonological Awareness (PIPA)	Pre-K, K and 1									✓					
Preschool Language Scale - 4th Edition (PLS-4)	Pre-K, K, and 1		✓												

Norm Reference Assessments	Grades	Cognitive Elements Evaluated by this Assessment													
		Reading Comprehension	Language Comprehension	Background Knowledge	Linguistic Knowledge	Phonology	Syntax	Semantics	Decoding	Cipher Knowledge	Lexical Knowledge	Phoneme awareness	Alphabetic Principle	Letter Knowledge	Concepts about print
Qualitative Reading Inventory --4th Edition (QRI-4)	K, 1, 2, 3, and higher	✓		✓					✓						
Rapid Automatized Naming and Rapid Alternating Stimulus Tests (RAN/RAS)	K, 1, 2, 3, and higher													✓	
Ready to Learn: A Dyslexia Screener	Pre-K, K, 1					✓		✓						✓	
Receptive One-Word Picture Vocabulary Test (ROWPVT)	Pre-K, K, 1, 2, 3, and higher							✓							
School Readiness Test (SRT)	K and 1 (Early 1st grade only)		✓					✓						✓	
Standardized Reading Inventory - 2nd Edition (SRI-2)	Pre-K, K, 1, 2, 3, and higher	✓						✓	✓						
Stanford Achievement Test -- 10th Edition (SAT-10)	K, 1, 2, 3, and higher	✓	✓					✓	✓	✓					
Stanford Diagnostic Reading Test -- 4th Edition (SDRT-4)	K, 1, 2, 3, and higher		✓					✓							
Stanford English / Spanish Language Proficiency Test (Stanford ELP / SLP)	K, 1, 2, 3, and higher	✓	✓												
Stanford Reading First	K, 1, 2, 3, and higher	✓	✓					✓	✓						
STAR Early Literacy Computer-Adaptive Diagnostic Assessment	Pre-K, K, 1, 2, and 3	✓		✓				✓	✓	✓			✓	✓	✓
STAR Reading Computer-Adaptive Reading Test	1, 2, 3 and higher	✓													
Test of Early Language Development - Third Edition (TELD-3)	Pre-K, K, 1, and 2		✓												
Test of Early Reading Ability-3rd Edition (TERA-3)	Pre-K, K, 1, 2, 3													✓	✓
Test of Kindergarten/First Grade Readiness Skills	K and 1	✓							✓					✓	
Test of Kindergarten/First Grade Readiness Skills (TKFGRS)	K and 1														
Test of Oral Reading and Comprehension Skills (TORCS)	K, 1, 2, 3, and higher	✓							✓						
Test of Phonological Awareness Skills (TOPAS)	K, 1, 2, 3, and higher											✓			
Test of Phonological Skills in Spanish (TPAS)	K, 1, 2, 3, and higher											✓			
Test of Reading Comprehension -- 3rd Edition (TORC-3)	2 and 3	✓					✓	✓							
Test of Silent Word Reading Fluency (TOSWRF)	1, 2, 3, and higher								✓						
Test of Word Reading Efficiency (TOWRE)	1, 2, 3, and higher								✓	✓					
Wechsler Individual Achievement Test - Second Edition (WIAT-II)	Pre-K, K, 1, 2, 3, and higher	✓	✓						✓	✓					

7.2. Summary of fluency-comprehension correlation studies

Authors	Year	Subjects	Criterion measure	Correlation
Fuchs, 1981; see also Fuchs & Deno, 1992	1981	91 1st to 6th grade with 15 special ed. & 23 Title 1	Woodcock Reading Mastery Test (WRMT)	.91 (Ginn) .91 (S-F)

Authors	Year	Subjects	Criterion measure	Correlation
Deno, Mirkin, and Chiang, 1982; see also Fuchs, Tindal, and Deno, 1984.	1982	45 1st to 6th grade with 18 LD	Cloze	.86 (3rd) .87 (6th)
Marston & Deno, 1982	1982	26 3rd grade general ed. 64 4th to 6th grade using lowest readers & Title 1	Questions below grade level Stanford Achievement Test (SAT) Science Research Associates (SRA)	.40 .88 .80
			Ginn 720 Reading Series Mazes	.83 (3rd) .76 (4th) .76 (5th) .59 (6th)
Buehner, 1983	1983	92 4th grade general ed.	Comprehensive Tests of Basic Skills (CTBS)	0.63
Fuchs, Tindal, Fuchs et al., 1983	1983	21 4th grade general ed.	Holt Basic Reading Series	0.79
Fuchs, Tindal, Shinn, Fuchs, and Deno, 1983; see also Tindal, Fuchs et al., 1985	1983	47 5th grade general ed.	Ginn Basal Reader Mastery Test	0.72
Lomax. 1983	1983	101 1st to 6th grade LD	Diagnostic Reading Scales (DRS)	0.76
			Comprehensive Tests of Basic Skills (CTBS)	0.75
Tindal, Fuchs et al., 1983; see also Tindal, Fuchs et al., 1985	1983	25 4th grade general ed.	Scott-Foresman Basal Mastery Test	0.7
Tindal, Shinn et al., 1983; see also Tindal, Fuchs et al., 1985	1983	47 6th grade general ed.	Houghton-Mifflin Basic Reading Test	0.66
Fuchs, Fuchs, and Maxwell, 1988	1988	35 4th to 8th grade with 27 LD, 7 ED, & 1 MR	SAT	0.91
			Questions	0.84
			Cloze	0.75
			Retell	0.74
Collins, 1989	1989	58 2nd grade general ed.	California Achievement Test (CAT) & Harcourt-Brace-Jovanovich (HBJ)	.75
			Basal Reader end of level test for comprehension	.60
			Word Meaning	.57 (3rd) .56 (6th)
Markell, 1991	1991	42 3rd grade general ed.	Mazes below grade level Mazes at grade level	.86 .89
Bain & Garlock, 1992	1992	479 1st to 3rd grade (1st was Title 1)	Comprehensive Tests of Basic Skills (CTBS)	.62 (1st) .79 (2nd) .72 (3rd)
Shinn, Good, Knutson, Tilly, and Collins, 1992	1992	238 3rd & 5th grades with 5% special ed.	Cloze SDRT Retell	.77 (3rd) .63 (5th) .59 (3rd) .58 (5th)

Authors	Year	Subjects	Criterion measure	Correlation
				.60 (3rd) .39 (5th)
Jenkins & Jewell, 1993	1993	335 2nd to 6th grade with 15 LD, 1 MR, 1 ED	Gates-MacGinite Reading Test (GMRT)	.86(2nd) .82(3rd) .86(4th) .68(5th) .63(6th)
			Metropolitan Achievement Test (MAT)	.84(2nd) .67(3rd) .82(4th) .64(5th) .58(6th)
Hintze, Shapiro, Conte, and Basile, 1997	1997	57 2nd to 4th grade with 8 special ed.	Degrees of Reading Power Test (DRP)	.67 (Auth) .64 (Liter.)
Kranzler, Brownell, and Miller, 1998	1998	57 4th grade general ed.	Kaufman Test of Educational Achievement-Brief (KTEA -B)	0.41
Madelaine & Wheldall, 1998	1998	50 1st to 5th grade general ed	Neale Analysis of Reading-Revised	0.71
Kranzler, Miller, and Jordan, 1999	1999	326 2nd to 5th grade general ed.	California Achievement Test (CAT)	.63 (2nd) .52 (3rd) .54 (4th) .51 (5th)
			Mazes above grade level	0.87
			Questions below grade level	0.26
			Questions at grade level	0.5
Jenkins, Fuchs et al., 2000	2000	113 4th grade with 7 RD	Iowa Test of Basic Skills (ITBS)	0.83

Source: Hosp, M. K. (2001). Variables that Affect the Correlation between Fluency and Accuracy with a Measure of Reading Comprehension. Nashville, TN: Peabody College of Vanderbilt University.

7.3. Random selection of schools, classrooms and students

Within each group of schools (full-grade, multigrade, and single-teacher), schools were ordered by province and size, so that there would be proper representation of each province and smaller schools would also have a chance to be selected. However, schools registering less than 10 students or located more than two days away from the district capital (by the most appropriate form of transportation) were not included in the sample because of the imbalance between the high cost of reaching them and the limited benefits of inclusion to total the sample.[50]

In each subgroup of schools, intervals of equal size were defined using the cumulative number of students. The starting point for the first interval was chosen randomly, and the schools that had the first student of each interval were selected. If any of the selected schools turned out to have too few students or to be located too far away (according to the aforementioned criteria), then the next school in the list was selected instead.

50 Schools in the districts of Pangoa and Mazamari were excluded from the sample for security reasons, as the area is currently under heavy military surveillance due to terrorist activity and drug trafficking. These districts account jointly for approximately 5.1 percent of the population of Junin (PNUD, 2006).

The selection of classrooms or sections was done by a random draw, typically in the presence of the principal and all grade teachers. The selection of students was done using the enrollment list to define the size of sampling intervals and randomly selecting the starting point. In third grade, ten students were selected randomly to take the oral test, which was administered individually outside class, but all students participated in the written test, as it required group administration inside the classroom. If the school had fewer than 10 students, the indication was to evaluate all third grade students present at the time.[51]

7.4. Population and sample description

a. Population description: Sample framework

Table 16. Distribution of primary public schools for minors by UGEL, type and location

UGEL	Single-Teacher		Multigrade				Full grade					Total
	Rural	Urban	Rural	Shanty town	Peri-urban	Urban	Rural	Shanty town	Peri-urban	Urban	Sub-urban	
Chanchamayo	217		139	3	3	4	5	3	4	6		**384**
Chupaca	9		55			2	9			6		**81**
Concepción	45		67		1	8	3			5	1	**130**
Huancayo	62		129	1	1	7	32	2	11	11	10	**266**
Jauja	16		77		8	17	6		1	13		**138**
Junín	39		19		5	1	1		3	4		**72**
Satipo	210		177		3	1	4		3	4		**402**
Tarma	76		69		2	1	2		3	17		**170**
Yauli	12	1	12			1	4		3	8		**41**
Total	**686**	**1**	**744**	**4**	**23**	**42**	**66**	**5**	**28**	**74**	**11**	**1.684**

Source: Padrón Escolar 2006.

Table 17. Distribution of third grade students of public primary schools for minors by UGEL and location in Junin, Peru

UGEL	Single-Teacher			Multigrade			Full grade			Total
	Rural	Urban	Peri-urban	Rural	Urban	Peri-urban	Rural	Urban	Peri-urban	
Chanchamayo	891	1		1,769	324	114	269	898	836	**5,102**
Chupaca	26	0		784	59		370	330		**1,569**
Concepcion	186			878	193	14	80	310	100	**1,761**
Huancayo	244	28	10	1,890	547	198	2,096	2,815	3,559	**11,387**
Jauja	43	0		914	301	98	229	949	35	**2,569**
Junin	104			171	40	85	19	323	98	**840**
Satipo	1,373	2	31	2,599	230	101	153	738	292	**5,519**
Tarma	185	0		869	168	41	79	1,289	138	**2,769**
Yauli	32	1		152	43		190	610	156	**1,184**
Total	**3,084**	**32**	**41**	**10,026**	**1,905**	**651**	**3,485**	**8,262**	**5,214**	**32,700**

51 Schools were given a quota of 10 students because we are interested in understanding the variance across schools and not within schools.

Table 18. Distribution of third grade students of public primary schools for minors by UGEL and gender in Junin, Peru

UGEL	Boys	Girls	Percent. Girls
Chanchamayo	2,628	2,474	**48.5%**
Chupaca	796	773	**49.3%**
Concepcion	875	886	**50.3%**
Huancayo	5,739	5,648	**49.6%**
Jauja	1,282	1,287	**50.1%**
Junin	412	428	**51.0%**
Satipo	2,794	2,725	**49.4%**
Tarma	1,354	1,415	**51.1%**
Yauli	606	578	**48.8%**
Total	**16,486**	**16,214**	**49.6%**

b. Sample description

For the psychometric analysis of the written test we use Sample A, which consists of all 1,001 students who took the written test. For the comparative analysis between the two tests, we use Sample B, which is a subsample of A that includes all 475 students who took both the oral and written tests. In both cases, the samples maintain the random selection from the original study.

Table 19. Schools in the sample by UGEL and type in the Sample

UGEL	Single-Teacher	Multigrade	Full grade	Total	%
Chanchamayo	9	7	4	**20**	**24**
Chupaca	0	1	2	**3**	**4**
Concepción	2	0	2	**4**	**5**
Huancayo	3	0	13	**16**	**19**
Jauja	0	1	3	**4**	**5**
Junín	2	0	1	**3**	**4**
Satipo	10	10	4	**24**	**29**
Tarma	3	2	2	**7**	**8**
Yauli	0	0	2	**2**	**2**
Total	**29**	**21**	**33**	**83**	**100**

Table 20. Distribution of third grade students by UGEL in Samples A and Sample B, compared to the third grade student population in Junin

	Population		Sample A (N=1,001)		Sample B (N=475)	
UGEL	**Freq.**	**Pct.**	**Freq.**	**Pct.**	**Freq.**	**Pct.**
Chanchamayo	5,102	15.60	155	15.48	85	17.89
Chupaca	1,569	4.80	38	3.80	20	4.21
Concepción	1,761	5.39	55	5.49	29	6.11
Huancayo	11,387	34.82	314	31.37	129	27.16
Jauja	2,569	7.86	71	7.09	34	7.16
Junín	840	2.57	20	2.00	11	2.32
Satipo	5,519	16.88	235	23.48	126	26.53
Tarma	2,769	8.47	64	6.39	25	5.26
Yauli	1,184	3.62	49	4.90	16	3.37
Total	**32,700**	**100.00**	**1,001**	**100.00**	**475**	**100.00**

c. Adjustment factor

Although the initial sampling approach allows us to run separate analysis for each type of school, when running statistical analyses for the sample as a whole it is necessary to correct the biases introduced with the sampling and to adjust the sample according to the real probability of selection. This was done using an adjustment factor (AF), calculated from a representation factor (RF) for each sample.

Table 21. Representation and Adjustment Factors by type of school

		Single-Teacher	Multigrade	Full grade	Total
Number of students	Sample A	120	285	596	**1,001**
	Sample B	104	164	207	**475**
	Population	3,157	12,582	16,961	**32,700**
Percentage of Students	Sample A	12%	28%	60%	**100%**
	Sample B	22%	35%	44%	**100%**
	Population	10%	38%	52%	**100%**
Representation Factor	Sample A	26	44	28	**99**
	Sample B	30	77	82	**189**
Adjustment Factor	Sample A	**27%**	**45%**	**29%**	**100%**
	Sample B	**16%**	**41%**	**43%**	**100%**

Since students were divided into three groups based on the type of school they attend for the selection process, the RF of each student in the sample is equivalent to the total number of students in his or her group in the population, divided by the total number of students in his or her group in the sample. The AF is calculated from the RF of each group, as a percentage of total RF, which is the sum of the RFs of all three groups (see Table 21). For example, the RF of Sample A for students in single-teacher schools is 3,157 (the total number of students in single-teacher schools in the sample) divided by 120 (the total number of students in single-teacher schools in the population), which results in 26. Then, the AF is 26 (the RF of single-teacher schools) divided by 99 (the sum of all RFs in this sample), multiplied by 100 (to have it as percentage).

7.5. Psychometric analysis of the ECE test with Raw vs. Transformed Rasch scores

To identify the trade-offs of using raw vs. transformed scores, two types of psychometric analyses were run: one for the items and one for the test as a whole. For evaluating the items we used the classic item analysis with raw scores, and the analysis under the Rasch model with transformed scores. The former will be referred to as analysis for raw scores and the latter as analysis for Rasch scores. For evaluating the tests, we used unidimensionality, reliability and normality analysis. This analysis was done with sample A, which includes 1001 cases, using the adjustment factor to match the proportion of students in each type of school to the proportions found in the population.

a. Calculating Raw and Rasch scores

For the written test, two sets of scores are calculated: raw and transformed. The raw total score is the sum of correct answers, that is, one point for each question answered correctly. The test contains four pairs of connected questions, namely items 2 and 3, 4 and 5, 6 and 7, and 8 and 9.[52]

[52] These items are two questions with the same three alternatives of answer, which means that answering one question reduces the chances of error on the other item.

Each of these pairs is treated as one item, the same way the MINEDU did for ECE (assigning one point if at least one of the items in the pair is answered correctly). Thus, the test comprises 23 questions but only 19 items, and the score range goes from zero to 19.[53]

The transformed scores were calculated using the Rasch model, which is what the MINEDU uses to analyze and report the results of its academic performance tests, and also the one used in international student performance assessments such as LLECE. Using a broader sample of 1,001 third-grade students who took the ECE test, we calculated three alternative scales of transformed scores using Rasch, setting the mean at 50 and created a scale from 0 to 100, with a standard deviation of 10. The first scale included all 1,001 cases and was calculated using WinBUGS software from a Bayesian perspective with the Markove chain Monte Carlo method (Kim, 2001), which estimates the abilities of all subjects, including extreme cases (those who answered all questions correctly or none).

MINEDU calculates ECE scores using the Rasch model[54] with Winstep, which automatically excludes the extreme cases because according to the authors of the estimation method these cases do not provide any information about the items (Wright & Stone, 1979). Thus, we calculated a second scale using this methodology, which included only 980 cases because 13 students scored 0, and 8 scored 19.

Finally, a common criticism to the use of multiple choice items in educational assessment is that there is no provision in the model for guessing, so a third scale was calculated using the Birnbaum's three-parameter logistic item response model to control for guessing (Baker & Kim, 2004; Birnbaum, 1968) but this model did not yield significant additional information. The estimations of the subjects' abilities obtained using this method were very similar to those obtained using the Rasch model, with a high correlation between the two ($r=0.9518$, $p<0.001$). The same was true for the estimations of the difficulty of the items ($r=0.966$, $p<0.001$). Furthermore, the guessing estimate, though positive for all items, is lower than the expected value of 0.33.

At the end, we decided to work with the first scale, which included all cases using the Rasch model, because when evaluating students for monitoring their progress, it is more important to have measures for all subjects than only for those who can provide information for the model (namely, the non extreme scores). The first scale was preferred also over the third one because it is a simpler model and provides similar information.

b. Item analysis

To evaluate the quality of individual items of the test we used both classic statistics and statistics based on the Rasch model using STATA and Winsteps. To evaluate the raw scores we used the percentage of right answers, the item-test and item-rest correlations, and the point-biserial correlations. To evaluate the Rasch scores we used the Infit and Outfit indicators and their z values.

For the raw scores, an item is considered to behave poorly from a psychometric point of view if at least two of the following statements are true: (i) the percentage of right answers is under 20 or over 80; (ii) the non-response rate is higher than 15 percent; (iii) the item-test or item-rest

53 There is also the possibility of transforming those scores to percentages of right answers, which would take values from 0 to 100, but we preferred to keep them in their raw form for simplicity and because transforming values to percentages does not provide any additional information.

54 For more information on Rasch modelling and its applications, see (Bond & Fox, 2007). Also see: (Wilson, 2004)

correlation is lower than 0.2 (validity indicator); or (iv) the point-biserial correlation is lower than 0.15 (discrimination indicator).

For the Rasch scores, an item behaves poorly if the Infit and Outfit values are lower than 0.9 or higher than 1.10; or the InZ and OuZ values are lower than -2 or higher than 2. An item should be eliminated if it behaves poorly in both models of analysis.

Table 22. Item Analysis Indicators for the ECE test (N=1001)

	Rasch Scores *					Raw Scores				Three-parameter Logistic:
Item (questions)	Difficulty **	IN. MSQ	IN. ZSTD	OUT. MS	OUT. ZSTD	Percent of correct response	Non-response	Item-rest correlation	Point-Biserial Correlation	Guessing probability ***
1 (1)	16.90	1.01	0.05	1.04	0.10	96.90	2.70	0.40	0.12	0.171
2 (2-3)	21.28	1.00	-0.03	1.32	0.95	96.00	3.40	0.39	0.16	0.170
3 (4-5)	20.44	0.97	-0.19	0.58	-1.58	96.20	2.60	0.37	0.17	0.160
4 (6-7)	20.00	1.05	0.29	0.88	-0.38	96.30	3.30	0.29	0.17	0.180
5 (8-9)	21.28	0.94	-0.36	0.66	-1.27	96.00	3.20	0.46	0.27	0.158
6 (10)	50.30	0.99	-0.16	1.02	0.25	71.80	5.80	0.42	0.22	0.170
7 (11)	49.26	0.97	-0.74	0.95	-0.81	73.50	4.60	0.43	0.22	0.186
8 (12)	58.11	0.98	-0.82	0.94	-1.57	57.30	6.30	0.40	0.27	0.156
9 (13)	67.74	1.10	3.63	1.13	2.80	37.90	6.70	0.47	0.27	0.183
10 (14)	62.77	1.06	2.40	1.03	0.76	47.90	6.30	0.30	0.24	0.177
11 (15)	60.14	0.91	-3.88	0.87	-3.55	53.20	6.70	0.31	0.25	0.126
12 (16)	46.89	0.99	-0.26	0.94	-0.74	77.10	5.60	0.29	0.25	0.161
13 (17)	44.71	0.90	-1.94	0.76	-3.09	80.10	6.10	0.15	0.11	0.177
14 (18)	60.82	0.92	-3.47	0.90	-2.89	51.80	7.40	0.20	0.16	0.134
15 (19)	67.18	1.01	0.54	1.08	1.82	39.00	6.60	0.36	0.34	0.206
16 (20)	68.47	1.02	0.75	1.07	1.45	36.50	7.20	0.30	0.23	0.227
17 (21)	69.52	1.03	1.00	1.11	1.99	34.50	7.60	0.39	0.33	0.228
18 (22)	71.55	1.05	1.59	1.15	2.47	30.80	7.20	0.34	0.32	0.171
19 (23)	72.64	1.06	1.64	1.14	2.23	28.90	7.60	0.22	0.19	0.158

* Statistics obtained in Winsteps for 981 cases, excluding the extreme cases (which answered all or non of the questions correctly).
** Mean of scale was set at 50.
*** Under the Three-Parameter Logistic Model (used to estimate guessing probability), an item is susceptible to guessing when the "guessing probability" is equal to or higher than the inverse of the number of alternative. In this case, since there are three alternatives, an item should not have a guessing probability equal to or higher than 0.33.

The results presented in Table 22 suggest that item 13 does no meet the criteria, as it systematically shows inadequate psychometric behavior both within and outside the Rasch model. We can also observe very high percentages of right answers in items 1 through 5 (equivalent to questions 1 through 9 in the test), as values range from 96 to 97 percent. This indicates that this set of items does not adequately discriminate reading ability for the sample because almost all children answer them correctly. Additionally, items 9, 11 and 14 (equivalent to questions 13, 15 and 18 in the test) do not adjust to the model.

Item 17 also presents a high percentage of right answers (80.1). Therefore, these items do not adequately discriminate reading comprehension in the sample. Table 23 shows the distribution of answers in each of the original questions, which confirms that the distracters in items 1 through 9 and 17 were not competitive enough against the right answer. Also, items 13, 19, 20, 21 and 22 show similar percentage of selection of each alternative (one third), as it is expected for random answers. Finally, question 23 has one distracter that has been chosen more often than the right answer.

In conclusion, the item analysis shows that items 13, 22 and 23 do not meet the criteria for good psychometric behavior. The high percentage of right answers in items 1 through 9 and 17 indicate that these items cannot discriminate between children with adequate and inadequate reading comprehension levels in the sample. It is likely that right answers in items 13, 19, 20, 21 and 22 were due to random choice. Finally, item 23 has a distracter with higher proportion of response than the right answer. This indicates that there were problems either in the construction of the test or in the administration of it during this study. It would be useful to compare these results with those of the national census 2006 to confirm where the problem lies.

Table 23. Response distribution by Item for the ECE test (N=1001)

	Frequencies					Percentages				
Question	No response	a	b	c	key	No response	a	b	c	Percent of correct response
1	26	970	0	5	a	2.6	96.9	0.0	0.5	96.9
2	27	5	11	958	c	2.7	0.5	1.1	95.7	95.7
3	31	11	948	11	b	3.1	1.1	94.7	1.1	94.7
4	23	11	109	858	c	2.3	1.1	10.9	85.7	85.7
5	21	956	10	14	a	2.1	95.5	1.0	1.4	95.5
6	31	9	954	7	b	3.1	0.9	95.3	0.7	95.3
7	31	9	5	956	c	3.1	0.9	0.5	95.5	95.5
8	28	66	8	899	c	2.8	6.6	0.8	89.8	89.8
9	27	7	958	9	b	2.7	0.7	95.7	0.9	95.7
10	57	56	719	169	b	5.7	5.6	71.8	16.9	71.8
11	45	736	134	86	a	4.5	73.5	13.4	8.6	73.5
12	62	574	225	140	a	6.2	57.3	22.5	14.0	57.3
13	66	277	379	279	b	6.6	27.7	37.9	27.9	37.9
14	62	479	219	241	a	6.2	47.9	21.9	24.1	47.9
15	66	82	320	533	c	6.6	8.2	32.0	53.2	53.2
16	55	772	68	106	a	5.5	77.1	6.8	10.6	77.1
17	60	802	60	79	a	6.0	80.1	6.0	7.9	80.1
18	73	253	156	519	c	7.3	25.3	15.6	51.8	51.8
19	65	291	390	255	b	6.5	29.1	39.0	25.5	39.0
20	71	292	273	365	c	7.1	29.2	27.3	36.5	36.5
21	74	324	345	258	b	7.4	32.4	34.5	25.8	34.5
22	71	308	253	369	a	7.1	30.8	25.3	36.9	30.8
23	74	436	202	289	c	7.4	43.6	20.2	28.9	28.9

To sum up, there are six items that do not provide adequate information to discriminate reading abilities when reported in raw scores, and these items account for 10 of the 23 questions in the test. In contrast, there are only three items that do not adjust well to the Rasch model, though even then, they can be used –with caution– to calculate the model and take advantage of the additional information it provides.

c. Test analysis

The test analyses include the evaluation of the test's normality, reliability and unidimensionality. Results show that the distributions of both raw and Rasch scores are not normal, there are no significant differences in reliability using raw or Rasch scores, and the test is not unidimensional.

Normality

To evaluate normality we use the one sample Kolmogorov-Smirnov test.[55] Table 24 shows that in both cases –raw and Rasch–, distributions are not normal. Histograms and Normal Q-Q plots of both distributions indicate that the lack of normality is due mainly to the outliers, especially those scoring at the bottom of the distribution (see Figure 13 and Figure 14). Therefore, we run the normality test excluding outliers and we find that although the distributions are closer to normal, they remain non-normal, as shown in Table 24.

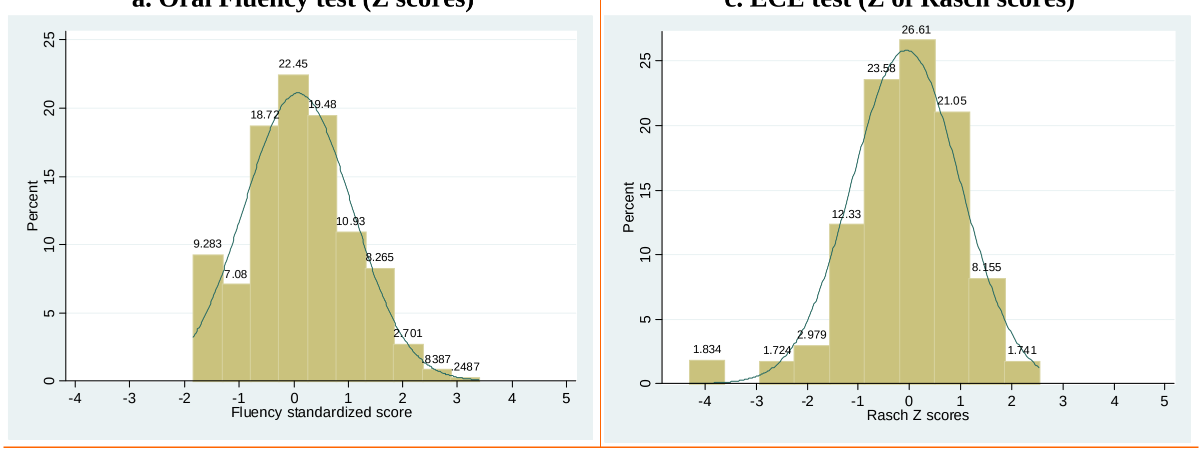
Table 24. Assessment of Normality for the Written Test (One-Sample Kolmogorov-Smirnov Test, N=1001)

		Sample A (N=1001)		Sample A excluding bottom outliers [c] (N=988)	
		Raw Scores	Rasch Scores	Raw Scores	Rasch Scores
Normal Parameters [a,b]	Mean	12.02	50.05	12.18	50.51
	Std. Deviation	3.006	8.10	2.69	7.09
Most Extreme Differences	Absolute	0.101	0.102	0.085	0.080
	Positive	0.065	0.071	0.081	0.080
	Negative	-0.101	-0.102	-0.085	-0.074
Kolmogorov-Smirnov Z		3.20	3.22	2.68	2.51
Asymp. Sig. (2-tailed)		0.000	0.000	0.0000	0.0000

[a] Test distribution is Normal.
[b] Calculated from data.
[c] The analysis excluded students with a score of 0 in the ECE test.

Figure 13. Histograms of Z scores in both tests (N=475)

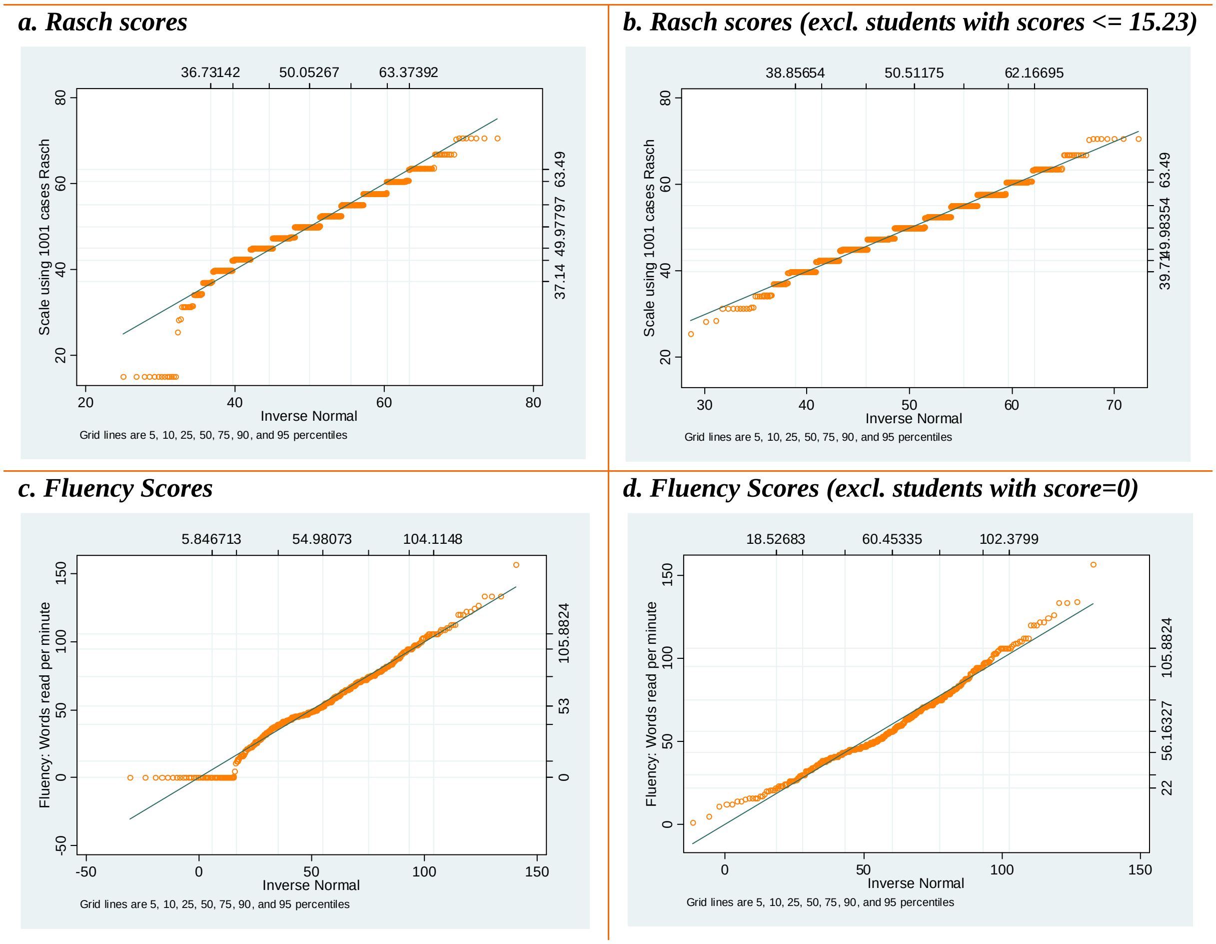
a. Oral Fluency test (Z scores) | **c. ECE test (Z of Rasch scores)**



---

55 The one-sample Kolmogorov–Smirnov test is a goodness of fit test used to compare the empirical distribution function with the cumulative distribution function specified by the null hypothesis. It is sensitive to any type of difference in the two distributions (shape, location, etc). The test is based on the largest difference between the two cumulative distributions.

Figure 14. Normal Q-Q Plots for Raw and Rasch Scores



Reliability

To test the reliability of the ECE we use the Cronbach's Alpha coefficient for the raw scores (Table 25).[56] For the Rasch scores, we use the Winsteps person reliability index (*pri*),[57] which is equivalent to the traditional "test" reliability of Cronbach's Alpha (Table 27). Results indicate that reliability of the test is similar for raw and Rasch scores, in both cases moderate: α=0.675 and *pri* = 0.61, respectively (excluding extreme cases for the latter). These coefficients are below the lowest bar recommended (0.7), suggesting that the reliability of the written test is not good enough. Given these results, one would expect that the internal problems of the test would also affect its correlation to other measures of the same or related constructs. However, internal reliability for Sample B (which is the one being used for comparison) is higher and above the

---

56 Alpha is computed by correlating all the scores on individual items, with the overall score on the test. Tests with high reliability, i.e. those with high internal consistency, will achieve an alpha coefficient of 0.75 or more on a scale of 0 to 1 where a high score indicates high reliability. The minimum alpha for acceptable reliability is 0.7.

57 The person reliability index "indicates the replicability of person ordering we could expect in this sample of persons were given another a parallel set of items measuring the same construct" (Bond & Fox, 2007, p. 40). The person reliability tells us if the test discriminates the sample into enough levels. Low values indicate a narrow range of person measures, or a small number of items. To increase person reliability, one can test persons with more extreme abilities (high and low) or lengthen the test. Improving the test targeting may help slightly. An index of 0.9 discriminates into 3 or 4 levels; an index of 0.8 discriminates into 2 or 3 levels; and an index of 0.5 discriminates into 1 or 2 levels. On the other hand, low item reliability means that the sample is not big enough to precisely locate the items on the latent variable. Literature supporting or elaborating on this index is very limited (see WINSTEPS).

minimum accepted level (α=0.74, see Table 26). When including extreme cases, the *pri* also goes up to 0.74 (see Table 28).

Table 25. Cronbach's Alpha coefficient for Sample A (N=1001)

Item (equivalent in the test)*	Sign	Item-test correlation	Item-rest correlation	Average inter-item covariance	alpha
I1 (Items 1+2)	+	0.389	0.338	0.018	0.664
I2 (Items 3+4)	+	0.365	0.306	0.018	0.665
I3 (Items 5+6)	+	0.431	0.377	0.018	0.661
I4 (Items 7+8)	+	0.378	0.323	0.018	0.664
I5 (Item 9)	+	0.438	0.383	0.017	0.661
I6 (Item 10)	+	0.421	0.287	0.016	0.660
I7 (Item 11)	+	0.439	0.309	0.016	0.657
I8 (Item 12)	+	0.439	0.292	0.016	0.659
I9 (Item 13)	+	0.310	0.154	0.017	0.677
I10 (Item 14)	+	0.364	0.208	0.017	0.670
I11 (Item 15)	+	0.495	0.353	0.016	0.651
I12 (Item 16)	+	0.427	0.303	0.017	0.658
I13 (Item 17)	+	0.500	0.391	0.016	0.649
I14 (Item 18)	+	0.485	0.342	0.016	0.652
I15 (Item 19)	+	0.377	0.225	0.017	0.668
I16 (Item 20)	+	0.366	0.216	0.017	0.669
I17 (Item 21)	+	0.353	0.204	0.017	0.670
I18 (Item 22)	+	0.326	0.179	0.017	0.673
I19 (Item 23)	+	0.327	0.183	0.017	0.672
Test scale**				**0.017**	**0.675**

* Items were re-numbered when calculating of a new total score over 19 that involved merging the pairs of connected items into single items.
** Test scale = mean(unstandardized items)

Table 26. Cronbach's Alpha coefficient for Sample B (N=475)

Item (equivalent in the test)*	Sign	Item-test correlation	Item-rest correlation	Average inter-item covariance	alpha
I1 (Items 1+2)	+	0.533	0.487	0.023	0.726
I2 (Items 3+4)	+	0.478	0.422	0.023	0.727
I3 (Items 5+6)	+	0.512	0.462	0.023	0.726
I4 (Items 7+8)	+	0.492	0.440	0.023	0.727
I5 (Item 9)	+	0.530	0.478	0.023	0.724
I6 (Item 10)	+	0.467	0.349	0.022	0.726
I7 (Item 11)	+	0.501	0.387	0.022	0.722
I8 (Item 12)	+	0.428	0.295	0.022	0.731
I9 (Item 13)	+	0.316	0.180	0.024	0.741
I10 (Item 14)	+	0.391	0.253	0.023	0.735
I11 (Item 15)	+	0.524	0.402	0.021	0.720
I12 (Item 16)	+	0.484	0.378	0.022	0.723
I13 (Item 17)	+	0.579	0.484	0.021	0.714
I14 (Item 18)	+	0.498	0.373	0.022	0.723
I15 (Item 19)	+	0.384	0.248	0.023	0.735
I16 (Item 20)	+	0.376	0.244	0.023	0.736

Table 26. Cronbach's Alpha coefficient for Sample B (N=475)

Item (equivalent in the test)*	Sign	Item-test correlation	Item-rest correlation	Average inter-item covariance	alpha
I17 (Item 21)	+	0.345	0.214	0.023	0.738
I18 (Item 22)	+	0.316	0.189	0.024	0.740
I19 (Item 23)	+	0.318	0.196	0.024	0.739
Test scale**				**0.023**	**0.740**

* Items were re-numbered when calculating of a new total score over 19 that involved merging the pairs of connected items into single items.
** Test scale = mean(unstandardized items)

The person reliability index (*pri*) under the assumptions of the Rasch model was obtained with the Winstep program excluding extreme cases. The *pri* is 0.61. Including extreme cases, reliability is 0.74.

Table 27. Summary of 980 measured (non-extreme) alums

REAL RMSE	6.53	ADJ.SD	7.47	Separation	1.14	Alum Reliability	.57
MODEL RMSE	6.20	ADJ.SD	7.75	Separation	1.25	Alum Reliability	.61
S.E. of alum mean = .32							

Table 28. Summary with 21 extreme alums= 1001 alums

REAL RMSE	7.01	ADJ.SD	11.15	Separation	1.59	Alum Reliability	.72
MODEL RMSE	6.71	ADJ.SD	11.33	Separation	1.69	Alum Reliability	.74

Maximum extreme score: 8 alums
minimum extreme score: 13 alums

Unidimensionality

Finally, unidimensionality is the most important premise for using the Rasch model. If a test is not unidimensional, the Rasch model should not be applied. To evaluate unidimensionality we use the Martin-Löf test in SAS (Christensen & Bjorner, 2003), factor analysis for dichotomic variables using the tetrachoric correlation matrix,[58] and Winstep's principal components of the correlation matrix of standardized residuals. The results of all these analyses show that the ECE test is not unidimensional. The factor analysis reveals five factors,[59] and the first factor explains less than 35 percent of the variance.[60]

The following table shows several analyses run to test the underlying unidimensionality in the 19 items of the ECE test. All the criteria used –for both raw and Rasch scores– point to the existence of several latent factors. The Scree plot, based on factor analysis of the tetrachoric correlation matrix indicates there are five factors with eigenvalues over 1, confirming the lack of unidimensionality of the test.

58 Orlando, Sherbourne and Thissen (2000) suggest that a set of questions can have multiple eigenvalues above 1, and therefore more than one factor, but still be unidimensional enough to be analyzed using Rasch. These authors argue that if there are enough items –say 80 percent– with loading over 0.35 in the first factor the test can be considered unidimensional. However, ECE the test does not meet any of these conditions in this sample.

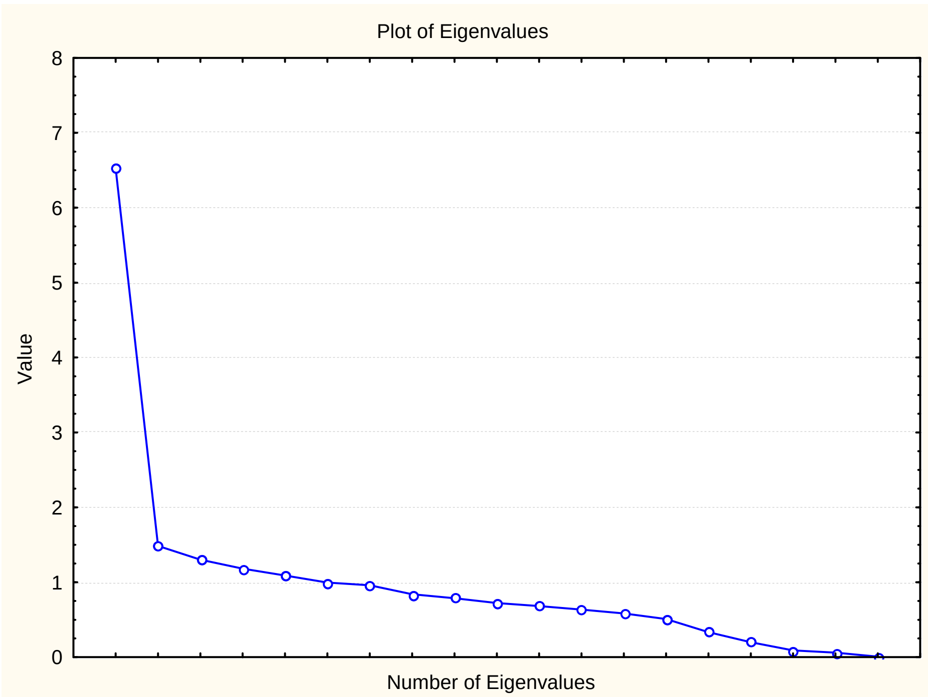
59 Using eigenvalues of 0.35 or more, the composition of the factors is as follows: Items 1 through 5 compose the first factor; the last three items (17-19) belong to the second factor; items 8, 10, 12 and 13 are part of the third factor; the fourth factor is formed by items 11, 14,15 and 16; and the fifth factor is composed by items 6, 7 and 9.

60 According to Carmines and Zeller (1979), a test can be considered unidimensional even if it has several factors as long as the first factor explains at least 40 percent of the variance.

Table 29. Results of the Unidinmentionality Analysis

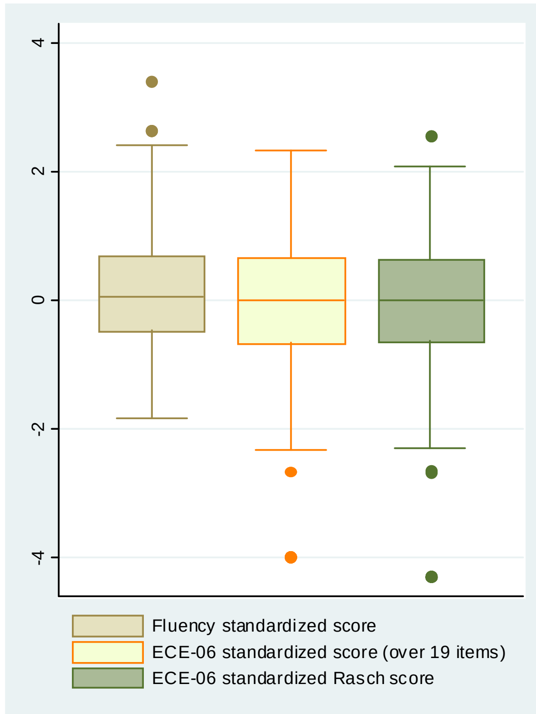
Method	Indicator	Score
Unidimensionality Test	Martin-Löf Test	292.88
	gl	89
	p	0.000**
Factor Analysis of dichotomic items using tetrachoric correlation matrix	Number of factors with eigenvalues over 1	5
	% variance explained by the first factor	34.35
	% cumulative variance	60.98
	% items with loading over 0.35 in the first factor	36.84
Principal Component analysis of standardized residual correlations for items under Rasch model	Number of factors with eigenvalues over 1	5
	% variance explained by the first factor	34.35
	% cumulative variance	60.98
	% items with loading over 0.35 in the first factor	36.84

Figure 15. Scree Plot based on factor analysis of the tetrachoric correlation matrix



7.6. Comparing Z-score distributions

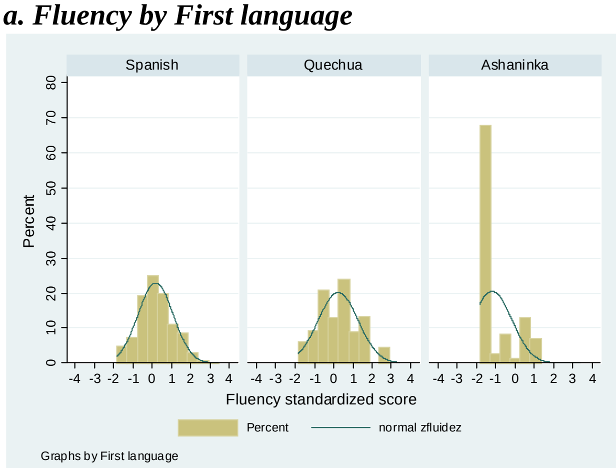
Figure 16. Box plots of Z scores in both tests (N=475, adjusted)



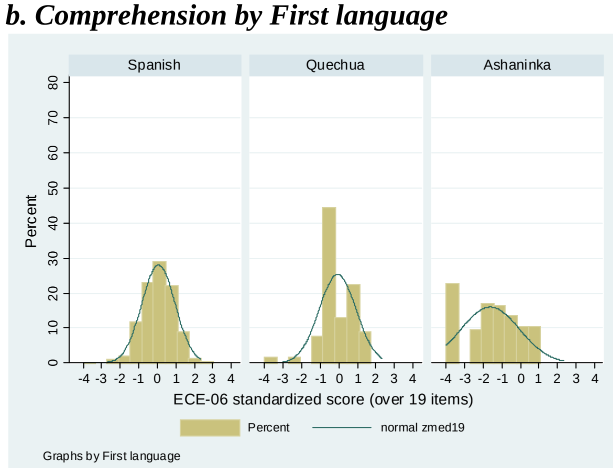
7.7. Comparison of distributions by groups

Figure 17. Distribution of Fluency and Comprehension Scores by First Language and Type of School

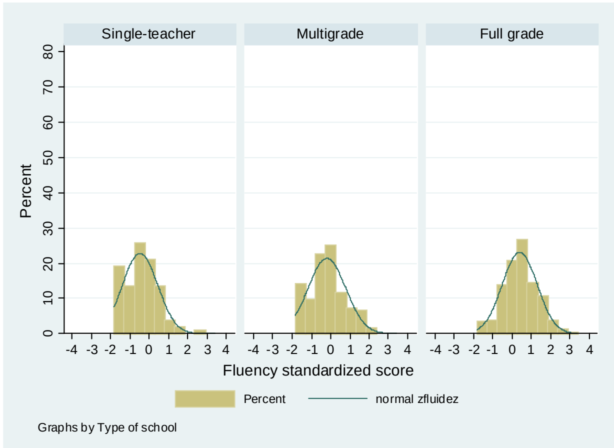
a. Fluency by First language



b. Comprehension by First language



c. Fluency by type of school



d. Comprehension by type of school

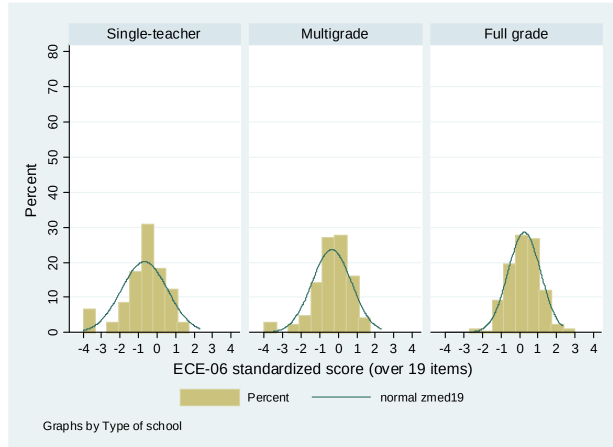


Figure 18. Box plots of Z scores in both tests by location of the school (N=475, adjusted)

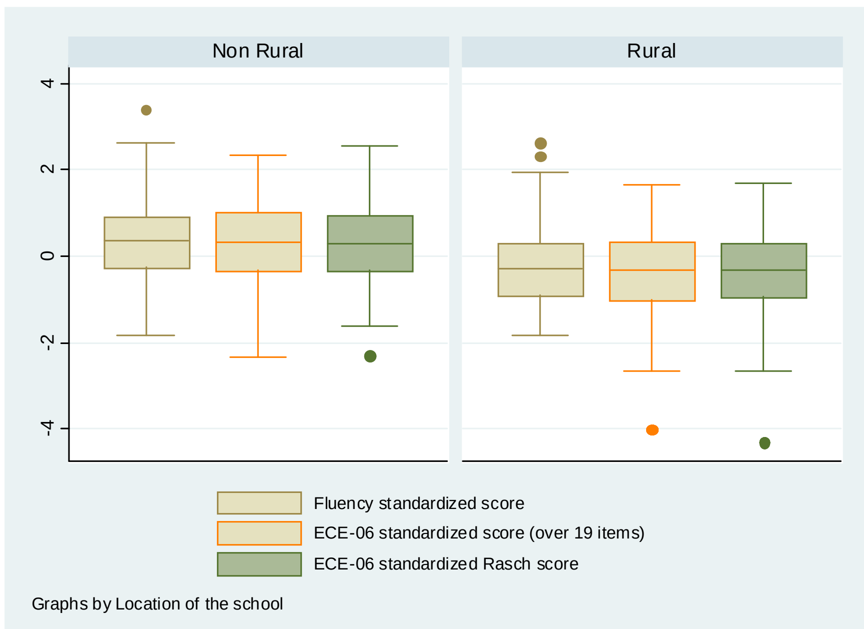


Figure 19. Box plots of Z scores in both tests by type of school (N=475, adjusted)

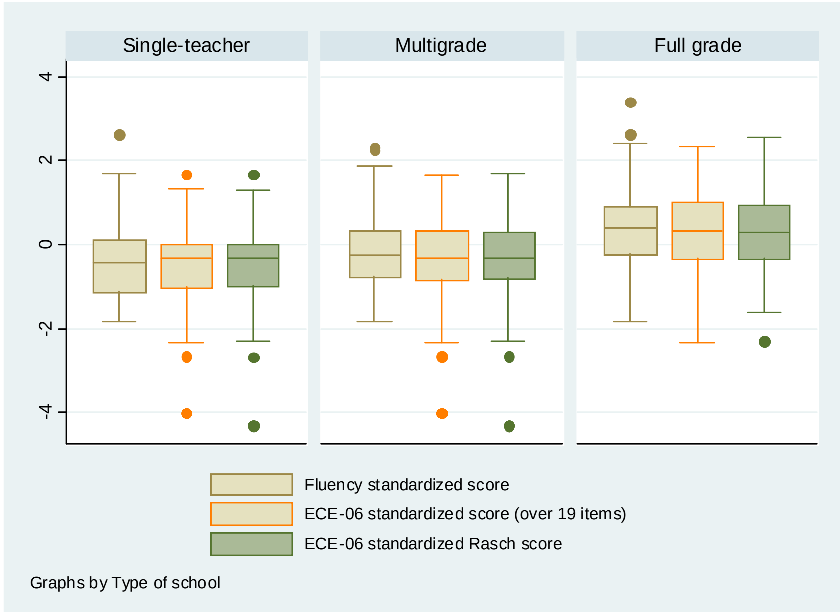


Figure 20. Box plots of Z scores in both tests by first language learned by the student (N=475, adjusted)

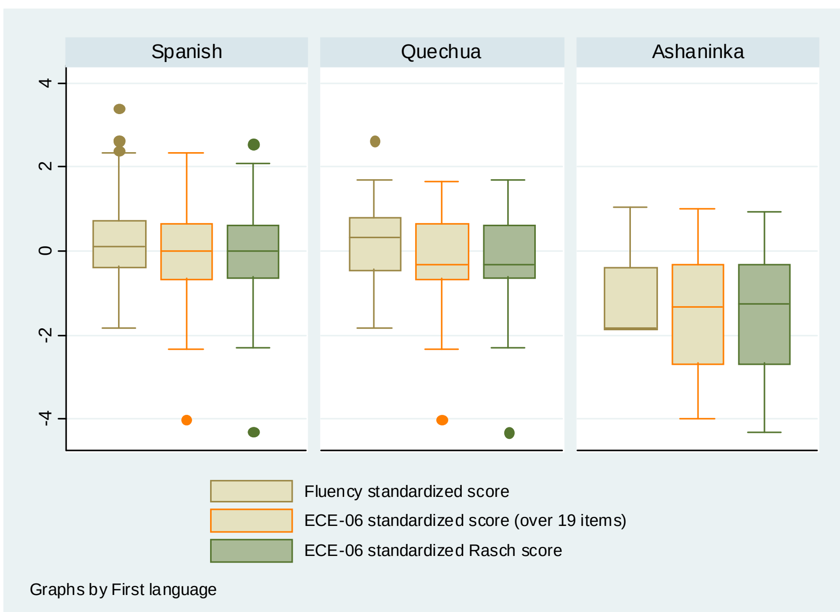


Figure 21. Box plots of Z scores in both tests by gender of the students (N=475, adjusted)

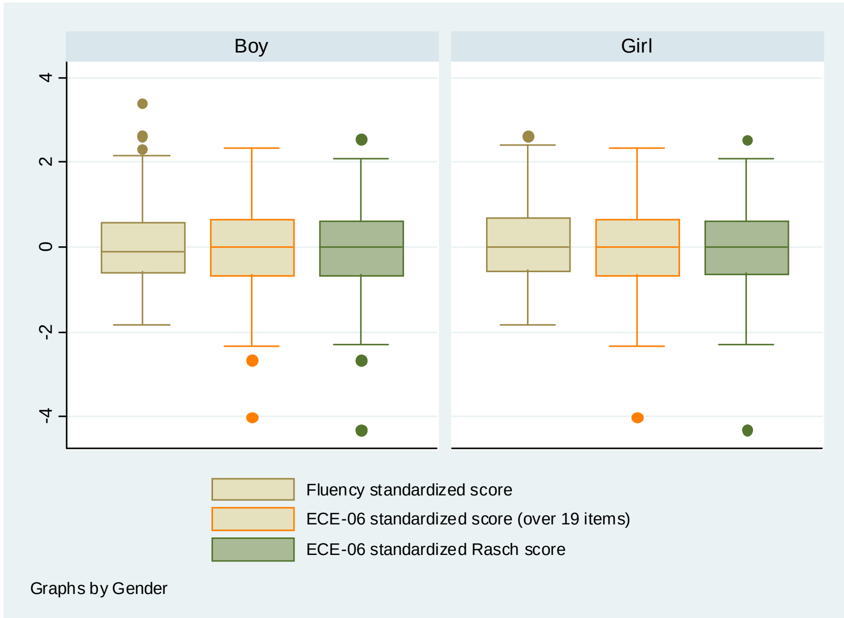
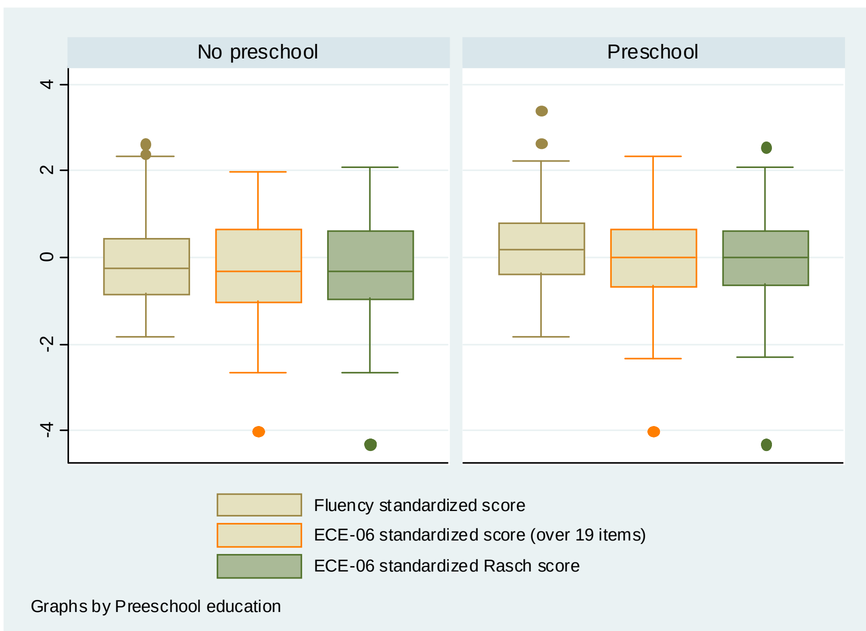


Figure 22. Box plots of Z scores in both tests by access to preschool education (N=475, adjusted)



7.8. Differential Item Functioning analysis: Cochran-Mantel-Haenszel Method

Table 30. Contingency table

Raw Score in written test	Non fluent (under 45 wpm)		Fluent (over 45 wpm)	
	Non Ashaninka	Ashaninka	Non Ashaninka	Ashaninka
0	3	9		
4	1	1		
5	4	2	1	
6	1	5	2	
7	3	3	4	
8	16	2	8	
9	13	3	19	1
10	13	3	40	2
11	26	2	31	
12	18	44	2	
13	13	43	1	
14	9	1	44	1
15	5	33	1	
16	3	19		
17	13			
18	1	4		
19	2			

Table 31. Cochran-Mantel-Haenszel

Fluency	Odds Ratio	95% Conf. Interval		M-H Weight	
Non fluent (under 45 wpm)	.0586667	.0097517	.2650043	7.075472	(exact)
Fluent (over 45 wpm)	.	0	.	0	(exact)
Crude	.0238095	.0040354	.1036947		(exact)
M-H combined	.0586667	.0147152	.2338923		
Test of homogeneity (B-D)		chi2(1) =	0.00		
		Pr>chi2 =	1.0000		
Test that combined OR = 1:		Mantel-Haenszel chi2(1) =	25.32		
		Pr>chi2 =	0.0000		

7.9. Item-test and Item-Rest correlations including Fluency as an item in the ECE test

Table 32. Item correlations including Fluency 45 wpm in the ECE test (N=475)

Item (equivalent in the test)*	Sign	Item-test correlation	Item-rest correlation	Average inter-item covariance	Alpha
I1 (Items 1+2)	+	0.528	0.485	0.024	0.744
I2 (Items 3+4)	+	0.479	0.427	0.024	0.745
I3 (Items 5+6)	+	0.517	0.471	0.024	0.744
I4 (Items 7+8)	+	0.495	0.448	0.025	0.745
I5 (Item 9)	+	0.527	0.478	0.024	0.743
I6 (Item 10)	+	0.461	0.350	0.023	0.744
I7 (Item 11)	+	0.494	0.386	0.023	0.741
I8 (Item 12)	+	0.431	0.307	0.023	0.748
I9 (Item 13)	+	0.323	0.196	0.025	0.757
I10 (Item 14)	+	0.390	0.261	0.024	0.752
I11 (Item 15)	+	0.530	0.417	0.022	0.739
I12 (Item 16)	+	0.476	0.376	0.023	0.743
I13 (Item 17)	+	0.575	0.486	0.022	0.734
I14 (Item 18)	+	0.500	0.383	0.023	0.742
I15 (Item 19)	+	0.374	0.246	0.024	0.753
I16 (Item 20)	+	0.369	0.244	0.024	0.753
I17 (Item 21)	+	0.339	0.216	0.024	0.755
I18 (Item 22)	+	0.303	0.183	0.025	0.757
I19 (Item 23)	+	0.308	0.193	0.025	0.756
Fluency 45 wpm	**+**	**0.514**	**0.406**	**0.023**	**0.740**
Test Scale				**0.024**	**0.757**

* Items were re-numbered when calculating of a new total score over 19 that involved merging the pairs of connected items into single items.
** Test scale = mean(unstandardized items)

7.10. Difference of means between fluent and less fluent readers

Table 33. Mean comparison at different cut points of fluency

Cut points of fluency	Mean and differences	Rasch score	Raw score
45 words per minute	Mean for 45 wpm or more	44.5	10.0
	Mean for less than 45 wpm	51.6	12.6
	Mean difference	**7.1**	**2.6**
60 words per minute	Mean for 60 wpm or more	46.0	10.5
	Mean for less than 60 wpm	52.5	12.9
	Mean difference	**6.5**	**2.4**
75 words per minute	Mean for 75 wpm or more	47.1	10.9
	Mean for less than 75 wpm	54.0	13.5
	Mean difference	**6.9**	**2.6**
Standard Deviation (all)		**7.8**	**2.9**

Table 34. Mean comparison with raw and Rasch Scores using 60 wpm cutpoint (U-Mann Whitney test)

	Raw scores			Rasch Scores		
	Obs	Rank Sum (mean)	Expected Rank Sum	Obs	Rank Sum (mean)	Expected Rank Sum
Under 60 wpm	273	53,775 (198)	64,974	273	53,938.5 (197)	64,974
Over 60 wpm	202	59,275 (293)	48,076	202	59,111.5 (293)	48,076
Combined	475	11,3050 (238)	113,050	475	113,050	113,050
Unadjusted variance			2187458			2187458
Adjustment for ties			-23551			-11.88
Adjusted variance			2163907			2187446
z			-7.502			-7.572
Prob > |z|			0.0000			0.0000

Table 35. Mean comparison with raw and Rasch Scores using 60 wpm cutpoint (U-Mann Whitney test)

	Raw scores			Rasch Scores		
	Obs	Rank Sum (mean)	Expected Rank Sum	Obs	Rank Sum (mean)	Expected Rank Sum
Under 60 wpm	361	76,284 (211)	85,918	361	36,766 (211)	85,918
Over 60 wpm	114	36,766 (323)	27,132	114	76,314 (322)	27,132
Combined	475	11,3050 (238)	113,050	475	11,3050 (238)	113,050
Unadjusted variance			1632442.00			1632442.00
Adjustment for ties			17575.49			-8.87
Adjusted variance			1614866.51			1632433.13
z			-7.581			-7.517
Prob > |z|			0.0000			0.0000

7.11. Rasch analysis with Fluency dummy variable in the ECE test (20 items)

Table 36. Scale Difficulty of ECE Items and Dichotomic versions of Fluency

	Analyses and Scale Difficulty						
ITEMS	**F30**	**F40**	**F45**	**F60**	**F75**	**F90**	**F110**
1	22.3	22.1	21.8	21.3	20.8	20.3	19.6
3	24.0	23.7	23.5	22.9	22.4	22.0	21.3
4	24.0	23.7	23.5	22.9	22.4	22.0	21.3
5	25.5	25.2	25.0	24.4	23.9	23.5	22.7
2	25.9	25.6	25.4	24.8	24.3	23.9	23.1
F30	**41.0**						
12	45.2	44.9	44.6	44.0	43.5	43.1	42.4
13	45.4	45.1	44.9	44.3	43.8	43.4	42.6
F40		**46.3**					
6	49.4	49.1	48.9	48.3	47.7	47.3	46.6
7	49.5	49.2	49.0	48.4	47.9	47.4	46.7
F45			**50.5**				
8	56.8	56.5	56.3	55.7	55.2	54.7	54.0
10	60.2	59.9	59.7	59.1	58.6	58.2	57.4
11	60.6	60.3	60.1	59.5	59.0	58.6	57.8
14	61.0	60.7	60.5	59.9	59.4	59.0	58.2
F60				**61.7**			
15	64.2	63.9	63.7	63.1	62.6	62.2	61.4
16	66.4	66.1	65.9	65.3	64.8	64.4	63.6
9	66.8	66.6	66.3	65.7	65.3	64.8	64.1
17	68.5	68.2	68.0	67.4	66.9	66.5	65.7
18	71.0	70.7	70.5	69.9	69.4	69.0	68.2
19	72.5	72.3	72.0	71.4	70.9	70.5	69.7
F75					**71.2**		
F90						**79.5**	
F110							**93.6**