EdData II

Análisis psicométrico de EGRA y su validez concurrente con otras evaluaciones de desempeño en lectura: caso Honduras y Nicaragua

15 diciembre 2010

EdData II Asistencia Técnica y Gerencia, Orden de Trabajo Número 3
Número de Contrato EHC-E-00-04-00004-00

Este informe se produjo para la Agencia de Los Estados Unidos para el Desarrollo Internacional. Fue elaborado por RTI International.

Análisis psicométrico de EGRA y su validez concurrente con otras evaluaciones de desempeño en lectura: caso Honduras y Nicaragua

Preparado para
Oficina de Desarrollo Económico, Agricultura y Comercio (EGAT/ED)
Agencia de los Estados Unidos para el Desarrollo Internacional (USAID)

Sometido por
Jorge Bazán
Pontificia Universidad Católica de Perú

Con el apoyo de
Amber Gove
RTI International
3040 Cornwallis Road
Post Office Box 12194
Research Triangle Park, NC 27709-2194

Índice

Página

Lista de figuras iv

Lista de cuadros v

Resumen ejecutivo vii

Executive summary xi

Introducción 1

Capítulo 1. Definición del estudio psicométrico de EGRA 3
1.1 Objetivos del estudio 3
1.2 Marco conceptual 3
1.2.1 Diagnóstico inicial de lectura (EGRA) 3
1.2.2 La evaluación psicométrica de un instrumento 13
1.3 Metodología del estudio 25
1.3.1 Definición de las muestras de estudio 25
1.3.2 Intrumentos considerados 26
1.3.3 Procedimientos para la evaluación psicométrica de EGRA 26

Capítulo 2. Resultados 27
2.1 Análisis de la confiabilidad de los aspectos medidos en EGRA 27
2.1.1 Resumen 30
2.2 Estructura de correlaciones de los aspectos medidos de EGRA 30
2.2.1 Caso Nicaragua 30
2.2.2 Caso Honduras 33
2.2.3 Resumen 35
2.3 Análisis de la estructura de EGRA 36
2.3.1 Caso Nicaragua 36
2.3.2 Honduras 41
2.3.3 Resumen 45
2.4 Análisis de la validez concurrente de EGRA 46
2.4.1 Nicaragua 46
2.4.2 Honduras 47
2.4.3 Resumen 48

Capítulo 3. Conclusiones y recomendaciones 51

Bibliografía 53

Anexo 1. Estadísticas descriptivas y correlaciones entre los aspectos medidos en EGRA en la versión de Nicaragua para el 2º, 3º y 4º grado y para la muestra completa 59

Anexo 2. Estadísticas descriptivas y correlaciones entre los aspectos medidos en EGRA en la versión de Honduras para el 2º, 3º y 4º grado y para la muestra completa 61

Lista de figuras

Página

Figura 1. Estructura de correlaciones entre los aspectos medidos de EGRA para el 2º, 3º y 4º grado en Nicaragua 32

Figura 2. Estructura de correlaciones entre los aspectos medidos de EGRA para el 2º, 3º y 4º grado en Honduras 34

Figura 3. Estructura de EGRA-Nicaragua considerando todos los subtest basados en un análisis factorial (solución de *minimum residual OLS*) con 3 factores con rotación *"oblimin"* usando la transformación Schmid-Leiman obtenida en *Psych Library* 39

Figura 4. Estructura de EGRA-Honduras considerando todos los subtest, basada en un análisis factorial (*minimum residual OLS solution*) con 3 factores con rotación *"oblimin"* usando la transformación Schmid-Leiman obtenida en *Psych Library* 44

Lista de cuadros

Página

Cuadro 1. Sistema de cuantificación de EGRA aspectos medidos 5

Cuadro 2. Muestras para la evaluación psicométrica de EGRA 25

Cuadro 3. Coeficientes Alfa de Cronbach para los diferentes subtest de EGRA en las versiones Nicaragua y Honduras 28

Cuadro 4. Alfa de Cronbach y otros índices de confiabilidad para la versión de EGRA Nicaragua (10 aspectos medidos y N=6649) 37

Cuadro 5. Cargas factoriales mayores que 0.2 en un análisis factorial con transformación Schmid-Leiman para los subtest de EGRA – Nicaragua 38

Cuadro 6. Índices de ajuste del análisis factorial confirmatorio de EGRA-Nicaragua para un modelo de tres factores y un modelo de un factor general (N=6649) 40

Cuadro 7. Alfa de Cronbach y otros índices de confiabilidad para la versión de EGRA-Honduras (10 aspectos medidos y N=1738) 41

Cuadro 8. Cargas factoriales mayors que 0.2 en un análisis factorial con transformación Schmid-Leiman para los subtest de EGRA-Honduras 42

Cuadro 9. Índices de ajuste del análisis factorial confirmatorio de EGRA-Nicaragua para un modelo de tres factores y un modelo de un factor general (N=1738) 45

Cuadro 10. Correlación de Pearson entre aspectos medidos en EGRA y puntaje fila y en escala para la prueba de español en una muestra de n=374 estudiantes de Nicaragua del 4º grado 46

Cuadro 11. Correlación de Pearson entre aspectos medidos en EGRA, puntaje fila y de escala de la prueba de español en estudiantes de Honduras (n=262 en 2º, n=213 en 3er, n=265 en 4º grado) 47

Resumen ejecutivo

El presente estudio es uno de los pocos exámenes, en el contexto de países en vías de desarrollo, que compara las propiedades de una prueba de lectura inicial, como el Diagnóstico de Lectura Inicial (Early Grade Reading Assessment, EGRA en inglés) con pruebas escritas orientadas a medir la comprensión, aplicadas en los mismos grados. Comparativamente, se han llevado a cabo repetidamente en países desarrollados estudios de correlación que permiten ciertas conclusiones en cuanto a la validez concurrente entre evaluaciones orales, orientadas a la fluidez, por un lado, y escritas, orientadas a la comprensión, por otro. Se ha encontrado que las correlaciones en estos casos de países desarrollados han sido razonablemente altas, por lo general 0.5 ó más. El estudio cubierto en este reporte encontró una fiabilidad interna entre moderada y alta para estos esfuerzos particulares de EGRA, y una correlación de alrededor de 0.4 entre la fluidez oral en un texto (lectura de un pasaje) y las medidas de comprensión en una evaluación escrita (prueba de desempeño en español).

Este estudio pudo sacar provecho del hecho de que en Nicaragua y Honduras se sometió al EGRA a muchos de los mismos niños que fueron evaluados mediante pruebas de comprensión escritas.

Para los datos utilizados en este estudio, en ambos países EGRA se aplicó en los grados 2º, 3º y 4º. La gama de habilidades tradicionales de EGRA fue evaluada en ambos casos. En Nicaragua, un subconjunto de 371 estudiantes fue sometido a una evaluación escrita de comprensión al final del 3er grado y luego, en 4º grado, fueron sometidos a una evaluación EGRA. En Honduras, 716 niños de los 3 grados fueron evaluados utilizando ambas herramientas. Por lo tanto, el tamaño de la muestra y la gama de los grados en cuestión fueron razonables. Todas las evaluaciones se llevaron a cabo en español. [1]

A partir de estos conjuntos de datos se puede derivar una impresión inicial de la co-validación y correlación entre estas evaluaciones orales y escritas. No se hace ningún juicio en cuanto a cuál debe servir de "ancla" o referencia. De este modo, el intento no es tanto evaluar la validez concurrente de EGRA con respecto a un ancla, sino evaluar su correlación y su capacidad de co-validarse entre sí. En la medida en que la comprensión de la lectura en silencio es una forma más "avanzada" de lectura, parece razonable, sin embargo, privilegiar un poco la evaluación escrita como ancla, sobre todo si ésta tiene características razonables como evaluación.

Antes que analizar la correlación y la validez concurrente de las dos evaluaciones, el informe analiza EGRA en sí mismo, sobre la base de una variedad de medidas tradicionales. Se halló que la fiabilidad entre las sub-pruebas del EGRA en sí era entre

[1] Como referencia, también téngase cuenta de que un tiempo antes de esta comparación, una evaluación doble similar—la cual produjo un conjunto de datos similares—tuvo lugar en Perú. Para detalles, ver Kudo y Bazan (2009).

moderada y alta, en general, sobre todo entre las que son las más comúnmente utilizadas como marcadores en general, por ejemplo, el reconocimiento de palabras familiares, el descifrar palabras desconocidas, y la fluidez en la lectura narrativa. Para estas sub-pruebas, el coeficiente Alfa fue de 0,8 a 0,9. En la mayoría de las aplicaciones, el nivel mínimamente adecuado de fiabilidad debe ser 0,7; entre 0,8 y 0,9 se considera moderadamente alto y 0,9 ó más se considera alto. Para algunas de las sub-pruebas menos importantes, como "orientación a la lectura" (dónde se comienza a leer, direccionalidad de lo impreso, etc.), se encontró que la fiabilidad no fue aceptable, en parte debido a que éstas mostraron tan poca variabilidad que fue difícil discernir si realmente se midió algo. Es importante señalar, al menos en Honduras y Nicaragua, donde los estudiantes han adquirido considerable habilidades de lectura oral en el 4 grado, la fiabilidad de los subtest de EGRA tiende a ser menor en este grado, en parte porque a los estudiantes les está yendo bastante bien (claro está que esto se produce por el menor grado de dificultad de esta aplicación particular del EGRA en este grado), por lo que es necesario tomar cuidado de las correlaciones entre EGRA y los niveles de desempeño en español de los estudiantes.

La fiabilidad de EGRA en su conjunto ha sido determinada por la estructura de correlación entre las sub-pruebas. Se encontró que la fiabilidad en ambos países es alta en la mayoría de los grados y cuando se incluyen la mayoría de las sub-pruebas, pero no todas. En general, como era de esperarse, las correlaciones entre las sub-pruebas que miden la fluidez fueron bastante altas. Las correlaciones entre la orientación a la lectura, la comprensión oral, y el dictado fueron bajas entre ellas, y también relativamente bajas en relación con las medidas de fluidez. La correlación entre las medidas de fluidez (todas ellas) y la comprensión fue buena, pero sólo en los primeros grados. Es decir, la correlación entre la fluidez y la comprensión fue menor en 4º grado. Más aún, en general, la estructura de las correlaciones se debilita a medida que avanzan los grados. Por lo tanto, los resultados confirman que EGRA es más apropiada para los primeros grados, en particular en los idiomas más fáciles y en los países donde la adquisición de la lectura es mayor que en los países de más bajo rendimiento. Es posible que una versión más exigente de la usual del EGRA pueda discriminar de manera útil en los grados más avanzados.

La coherencia interna de EGRA fue analizada mediante el cálculo de los valores Alfa y Omega de la evaluación en su conjunto. La consistencia interna no siempre fue alta cuando las sub-pruebas menos confiables, como la orientación a la lectura, fueron incluidas. Una vez que este tipo de sub-pruebas fueron excluidas, los valores Alfa y Omega fueron alrededor de 0,8 a 0,85, es decir, son moderadamente altos, indican una buena consistencia interna y proporcionan una sensación de que por lo menos para la mayoría (pero no todos) de los componentes hubo, efectivamente, un concepto unificado subyacente que puede confiablemente llamarse "lectura inicial."

Al evaluar la validez concurrente de EGRA y las pruebas escritas (o, más precisamente, la correlación entre la sub-prueba clave en EGRA—fluidez oral en lectura narrativa

(lectura en un pasaje)—y el puntaje general en la prueba de lápiz y papel (prueba de desempeño en español)), hay que señalar que ninguna de ellas fue diseñada específicamente para ser comparada con otra. Sin embargo, en cuanto a la sub-prueba clave (la fluidez en lectura de textos narrativos o lectura de un pasaje), las correlaciones bordean un promedio de entre 0,35 y 0,4, cuando los resultados del 2º grado en Honduras, que parecen ser un caso anómalo, se excluyen. Un estudio más interesante consistiría en comparar una sub-prueba clave de EGRA, como la fluidez oral en textos narrativo o lectura de un pasaje, con los ítems o elementos individuales de una prueba de comprensión escrita y no solo con el puntaje general de una prueba de desempeño en español. Esto no fue posible con los datos de Nicaragua y Honduras. Una evaluación de la correlación de la fluidez y las evaluaciones escritas que fueron específicamente diseñadas para esto (pero que no fue parte de este estudio, aunque esté aquí reportado) demostraron una correlación entre la fluidez y la comprensión, de alrededor de 0,5. El estudio, curiosamente, muestra que la correlación entre la fluidez y la evaluación general en comprensión fue tan buena como la correlación promedio entre cualquiera de los ítems de comprensión y la evaluación general en comprensión (ver Kudo y Bazán, 2009).

La conclusión de este estudio es que una evaluación enfocada en las capacidades orales relacionadas con la fluidez es de moderada a altamente confiable (dependiendo de la sub-prueba de que se trate; en general, las de fluidez son más fiables) y tiene una correlación moderada con las medidas de comprensión. Este hallazgo, junto con la evidencia proveniente de los países desarrollados, es alentador en relación con la aplicación de las evaluaciones orales, como un precursor o un proxy para las habilidades más avanzadas. Se deben de llevar a cabo estudios adicionales sobre las propiedades de estas evaluaciones, tanto para mejorarlos como para aumentar la facilidad de su uso, utilizando métodos que estén diseñados específicamente para este propósito.

Executive summary

The present study is one of the few examinations—in the context of developing countries—comparing the properties of an early grade oral assessment such as the Early Grade Reading Assessment (EGRA) to written, comprehension-oriented assessments applied to the same grades. By contrast, correlation studies have been done repeatedly in developed countries, thereby enabling some claims as to concurrent validity between oral, fluency-oriented assessments and written, comprehension-oriented assessments. The correlations in these developed-country cases have been found to be reasonably high—usually 0.5 or more. The study covered in this report found moderate to high internal reliability of these particular EGRA efforts, and a correlation of around 0.4 between oral fluency (connected text fluency) in connected text and comprehension measures in a written assessment (achievement test in Spanish).

This study was able to take advantage of the fact that in Nicaragua and Honduras, EGRA was administered to many of the same children who had been assessed using pencil-and-paper comprehension tests. For the data used in this study, in both countries, EGRA was applied in grades 2, 3, and 4. The traditional EGRA battery of skills was assessed in both cases. In Nicaragua, a subset of 371 students underwent a pencil-and-paper comprehension assessment in grade 3 and, later, in grade 4, underwent an EGRA assessment. In Honduras, 716 children drawn from all three grades were assessed using both tools. Thus, the sample sizes and the span of grades in question were reasonable. All assessments took place in Spanish.[2]

From these data sets, an initial impression of the co-validation and correlation between the oral and written assessments can be derived. No judgment is made as to which serves as "anchor." Thus, the attempt is not so much to assess the concurrent validity of EGRA with respect to an anchor, as to assess their correlation and their ability to co-validate each other. To the extent that comprehension in silent reading is a more "advanced" goal, however, it seems reasonable to somewhat privilege the pencil-and-paper assessment as an anchor, if it also has reasonable characteristics as an assessment.

Before analyzing the correlation and concurrent validity of the two assessments, the report analyzes EGRA itself, against a variety of traditional measures. Reliability *within* the EGRA subtests was found to be moderate to high, in general, especially among the EGRA subtests that are most commonly used as general markers, such as familiar word recognition, unfamiliar word decoding, and fluency in reading connected text. For these subtests, the alpha coefficient was around 0.8 to 0.9. In most applications, a minimum appropriate level of reliability is considered to be 0.7, 0.8 to 0.9 is considered moderately high, and 0.9 and above is considered high. Some of the less important subtests, such as orientation to print, were found to have unacceptable reliability; partly because they

[2] For reference, also note that a while before this comparison, a similar double assessment—producing a similar data set—took place in Peru. For details, see Kudo and Bazan (2009).

showed so little variability that it was hard to discern whether any measurement was taking place. It is important to note that, at least in Nicaragua and Honduras, where students have acquired considerable oral reading skill by grade 4, the reliability deteriorates for this grade, again, partly because students are all doing quite well (relative, that is, to the low level of difficulty of this particular EGRA application in this grade), so it is necessary take care to correlations between EGRA and levels of proficiency in Spanish across students.

The reliability of EGRA as a whole was determined by the structure of correlation *between* subtests. This correlation was found to be high in both countries, in most grades, and when including most of the subtests but not all. In general, as one would expect, the correlations between items measuring fluency were quite high. The correlations between orientation to print, listening comprehension, and dictation were low between themselves and also were relatively low with respect to the fluency measures. The correlation between fluency measures (all of them) and comprehension was good, but only in the earlier grades. That is, interestingly, the correlation between fluency and comprehension was lower in grade 4. Furthermore, in general, the structure of correlations weakened as the grades progressed. Thus, the results confirm that EGRA is more appropriate for the *early* grades, particularly in easier languages and in countries where reading acquisition is higher than in the lowest-performing countries. It is possible that a more demanding version of EGRA than has been customary could also usefully discriminate in later grades.

The internal coherence of EGRA was analyzed by calculating alpha and omega values for the assessment as a whole. The internal consistency was not always high if the less-reliable subtests—such as orientation to print—were included. Once these sorts of items were excluded, the alpha and omega values were around 0.8 to 0.85, which is moderately high, indicates good internal consistency, and provides a sense that at least for *most* (but not all) of the components there was an underlying construct that can be reliably called "early reading."

In assessing concurrent validity of EGRA and the written tests (or, more specifically, the correlation between the key EGRA subtest—fluency in reading connected text—and the overall score on the pencil-and-paper test (achievement test in Spanish), it has to be noted that neither effort was designed specifically to be compared to the other. With respect to the key subtest (fluency in reading connected text), the correlations were found to average 0.35 or 0.4 if the grade 2 results from Honduras, which seem to be an outlier, were excluded. A more interesting study would compare a key EGRA subtest such as fluency in connected text with the individual items in a pencil-and-paper comprehension test, not just with the overall score of a achievement test in Spanish. This was not possible using the Nicaragua and Honduras data. An assessment of the correlation of fluency and written assessments that was specifically designed for this purpose (but was not part of this study, though it is reported here) showed a correlation between fluency and comprehension of around 0.5. The study, interestingly, shows that the correlation

between fluency and the overall comprehension assessment was as good as the average correlation between any of the comprehension items and the overall comprehension assessment. (See Kudo and Bazan, 2009).

The conclusion of this one study is that an assessment focused on oral capabilities related to fluency is moderately to highly reliable (depending on the subtest in question, with the fluency ones generally being more reliable) and moderately well correlated with measures of comprehension. This finding, together with the evidence that exists from developed countries, is encouraging with respect to the application of oral assessments as a precursor or proxy for more advanced skills. Further study of the properties of these assessments, both to improve them and to increase their usability, should be conducted, using approaches that are specifically designed for this purpose.

Introducción

Con objetivo de ayudar al desarrollo de experiencias en proyectos especiales, USAID coordinó con RTI International la realización de un estudio sobre habilidades de lectura en Nicaragua, en los meses de abril y mayo 2008, con una muestra nacional de 126 escuelas seleccionadas aleatoriamente del universo de escuelas del país con un total de 6649 estudiantes de 2º, 3º y 4ro grado. El instrumento principal aplicado fue la versión en español de la prueba Diagnóstico de Lectura Inicial (Early Grade Reading Assessment, EGRA en inglés). Para mayores detalles del proceso de aplicación ver CIASES y RTI International (2009).

EGRA está diseñada para diagnosticar capacidades de lecto-escritura y consta de un conjunto de ocho secciones: ubicación espacial para leer un párrafo, reconocimiento de letras, dominio de fonemas, lectura de palabras simples, decodificación de palabras, fluidez de lectura, comprensión lectora, comprensión oral y capacidad de tomar dictado.

Dos características importantes y peculiares de EGRA son su naturaleza oral, y su aplicación individual. Para más información, por favor, consultar el *Manual para la evaluación inicial de la lectura en niños de educación primaria* (RTI International, 2009) disponible en www.eddataglobal.org.

Posteriormente, entre agosto y diciembre de 2008, un estudio similar, en el que se aplicó EGRA, fue financiado por el Banco Mundial y desarrollado en Honduras con 72 escuelas seleccionadas aleatoriamente de las escuelas participantes del proyecto PROHECO[3], que incluyeron también alumnos de 2º, 3º y 4º grado, en un total de 1738. Para mayores detalles de este proceso ver CIASES (2010).

En el caso de Nicaragua, algunos estudiantes de 4º grado que dieron la prueba EGRA entre abril y mayo de 2008 también habían participado en el *Estudio anual 2007: Rendimiento académico de los estudiantes de 3er grado en español y matemáticas en las escuelas validando el nuevo currículo* (Agencia de los Estados Unidos para el Desarrollo Internacional [USAID], EQUIP1, 2008) para el cual fueron seleccionados 2731 estudiantes de 62 escuelas de un universo de 100 que participaron del proceso de validación curricular en Nicaragua. De esta manera, se cuenta con una muestra de 374 estudiantes que dieron ambas pruebas: la prueba de español, cuando cursaban el tercer grado en diciembre de 2007 (final del año escolar), y EGRA, en mayo de 2008 (inicio de año escolar), cuando cursaban el 4º grado.

En el caso de Honduras, algunos de los estudiantes que dieron EGRA también participaron de la Evaluación de los Aprendizajes de Español y Matemática de 2008, un estudio que incluyó 854 escuelas de 1ro a 6to grado con un total de 101895 estudiantes.

[3] Para mayor información acerca de las escuelas PROHECO ver http://www.se.gob.hn/index.php?a=Webpage&url=PROHECO_home

De esta manera, se cuenta con una muestra de 716 estudiantes de 2º. 3º y 4º que dieron ambas pruebas, la prueba de español y EGRA, aplicada en Octubre 2008, al final del año escolar.

Las pruebas descritas en los dos párrafos anteriores son escritas y se aplican en grupo.

Es decir, se cuenta con un subconjunto de estudiantes que dieron ambas pruebas (las de EGRA y las de español) en sus respectivos países, lo que constituye muestras ad hoc para una importante evaluación: la validez concurrente de EGRA, sin prejuicio de cuál de las dos pruebas debe servir de "ancla" o referencia". A partir de estas muestras ad hoc se va a considerar una evaluación de los aspectos de correlación y co-validación entre las pruebas. Este trabajo junto al de Kudo y Bazán (2009) permite comenzar a conformar una acumulación de análisis que contribuye a construir conocimientos sobre los comportamientos *relativos* de pruebas orales de lectura inicial y pruebas de desempeño de lápiz y papel silentes.

Este documento presenta información acerca de los aspectos psicométricos de Early Grade Reading Assesment (EGRA) en su versión en español, aplicado en Honduras y Nicaragua.

El enfoque de la evaluación psicométrica de la prueba de EGRA se apoya en diversas fuentes y en la literatura citadas oportunamente, e incluye principalmente la evaluación de la confiabilidad de la prueba y de los subtest que la componen, así como una evaluación de la validez concurrente de EGRA en relación con las pruebas de español en Honduras y Nicaragua.

Para el logro de los resultados que serán presentados fue necesaria la preparación y/o adecuación de las bases de datos para realizar los análisis psicométricos pertinentes. Estos análisis fueron diseñados considerando los procesos y desempeños, grados escolares, subtest de EGRA y las pruebas de español consideradas y fueron realizados en software especializados, como se detallará posteriormente.

Este reporte se encuentra estructurado en tres capítulos. El primero explica los objetivos del estudio, su marco conceptual y metodología. El segundo presenta los principales resultados de la evaluación psicométrica de EGRA. Finalmente, en el tercer capítulo se extraen conclusiones acerca de los resultados más importantes para someter a discusión algunos temas clave, y se generan recomendaciones para mejorar la calidad educativa en el país.

Capítulo 1. Definición del estudio psicométrico de EGRA

En este capítulo presentamos los objetivos del estudio, el marco conceptual empleado y la metodología del estudio.

1.1 Objetivos del estudio

Como objetivo central se planteó evaluar las características psicométricas de EGRA y, de manera específica:

1. Evaluar la confiabilidad de cada uno de los subtests de EGRA.
2. Identificar la estructura correlacional de los subtests de EGRA.
3. Evaluar la validez concurrente de EGRA con pruebas escritas de desempeño en español.

1.2 Marco conceptual

1.2.1 Diagnóstico inicial de lectura (EGRA)

Historia de EGRA

El diseño de EGRA se inició en octubre de 2006, cuando USAID, a través de su proyecto EdData II, contrató a RTI International para diseñar un instrumento con el cual evaluar la lectura en los primeros grados. El objetivo era ayudar a los países socios de USAID a iniciar el proceso de medir cuán bien los niños de los grados iniciales de primaria vienen adquiriendo las habilidades de lectura y, en última instancia, estimular esfuerzos más efectivos para mejorar el desempeño de esta habilidad nuclear del aprendizaje.

A partir de una revisión de las investigaciones y de las herramientas y evaluaciones de lectura existentes, RTI diseñó un protocolo para una evaluación oral individual de habilidades básicas de lectura de los educandos. Para conseguir observaciones a este protocolo y confirmar la validez del enfoque global, RTI convocó una reunión de científicos cognitivos, expertos en la enseñanza de lectura en los primeros grados, y expertos en metodología de investigación y en evaluación, para que revisaran los componentes claves propuestos del instrumento. En el taller se encargó a los participantes que cubrieran la brecha existente entre investigación y práctica; esto es, que fusionaran los avances en los campos de la lectura e investigación de la ciencia cognitiva con prácticas de evaluación alrededor del mundo. Los investigadores y profesionales presentaron evidencias sobre la manera de medir la adquisición de la lectura en los primeros grados de primaria. Además se les pidió que identificaran los puntos clave que debían considerarse al diseñar un protocolo de evaluación de lectura en los primeros grados, para múltiples países y lenguas. El taller, organizado por USAID, el Banco Mundial y RTI en noviembre de 2006, incluyó a más de una docena de expertos de un

grupo diverso de países, así como unos 15 observadores de instituciones, tales como USAID, el Banco Mundial, la William and Flora Hewlett Foundation, la Universidad George Washington, el Ministerio de Educación de África del Sur y el Plan International, entre otros. Desde entonces, EGRA ha sido aplicado en más de 50 países e idiomas por instituciones, gobiernos, organizaciones no gubernamentales (ONGs) y escuelas. La aplicación en Nicaragua, en 2008, fue la primera a nivel nacional en América Latina. Honduras le siguió, el mismo año, en una muestra de sus escuelas rurales.

Descripción del instrumento

EGRA está diseñada para diagnosticar capacidades de lecto-escritura y consta de un conjunto de ocho secciones. En la versión de Nicaragua, las secciones son: empeño y relación con a la letra impresa, conocimiento del nombre de las letras, conciencia fonémica y fonológica, conocimiento de palabras simples, decodificación de palabras sin sentido, lectura y comprensión de un pasaje, comprensión oral y dictado. En la versión de Honduras, las secciones son: conocimiento del nombre de las letras, identificación del sonido inicial, conocimiento de los sonidos de las letras, lectura de palabras simples, lectura de palabras sin sentido, lectura y comprensión de un pasaje, comprensión oral y dictado.

Ambas versiones son similares, pero no iguales, (ambas tienen 8 secciones pero en la versión de Nicaragua se miden 10 aspectos y en la de Honduras 9). También existen diferencias menores en algunos enunciados y en los ejemplos, las instrucciones para los aplicadores, las formas de registro y el estímulo dado a los examinados en los aspectos medidos. Ambos instrumentos aparecen en los respectivos informes e resultados para ambos países, ya citados (ver CIASES y RTI International, 2009; CIASES 2010).

Nosotros vamos a ser referencia a los aspectos medidos antes que a las secciones, considerando además que no necesariamente los nombres de las secciones coinciden entre los países y que además una determinada sección permite más de un aspecto medido. Los aspectos medidos están basados además en el proceso de cuantificación llevado a cabo.

Los aspectos específicos que son medidos en al menos una versión de EGRA los cuales son presentados en el cuadro 1, son los siguientes:

- Ubicación espacial para leer un párrafo.
- Nombramiento de las letras.
- Identificación del sonido de las letras.
- Identificación del sonido de la letra inicial de una palabra.
- Discriminación del sonido inicial de palabras.
- Lectura de palabras simples.
- Decodificación de palabras sin sentido o “palabras inventadas”.
- Lectura de un pasaje.

Cuadro 1. Sistema de cuantificación de EGRA aspectos medidos

Aspecto medido	Habilidades	Elementos (ítems)	Formas de cuantificación	Nombre de la sección en Nicaragua	Nombre de la sección en Honduras	Name in English	Abr
Ubicación espacial para leer un párrafo		3	Porcentaje de respuestas correctas de 3 preguntas	1. Empeño y relación con la letra impresa	No aplicado	Reading direction	RD
Nombramiento de las letras	Fluidez y precisión de lectura de letras	100	Precisión o número de letras nombradas correctamente, velocidad o número de letras nombradas correctamente por minuto	2. Conocimiento del nombre de las letras	1. Conocimiento del nombre de las letras	Letter name recognition	LNR
Identificación del sonido de la letra inicial de una palabra	Conciencia fonética	10	Porcentaje de respuestas correctas de 10 preguntas	3. Conciencia fonética y fonológica	2. Identificación del sonido inicial	Letter sound recognition by Word	LSR1
Identificación de palabras que inician con el mismo sonido	Conciencia fonológica	10	Porcentaje de respuestas correctas de 10 preguntas	3. Consciencia fonética y fonológica	No aplicado	Letter sound recognition between word	LSR2
Recuerdo del sonido de las letras	Fónica de letras	50	Precisión o número de sonidos de letras recordados correctamente, velocidad o número de sonidos de letras recordados correctamente por minuto	No aplicado	3. Conocimiento del sonido de las letras	Letter sound recall	LSR3
Lectura de palabras simples	Fónica de palabras	50	Precisión o número de palabras leídas correctamente, velocidad o número de palabras leídas correctamente por minuto	4. Conocimientos de palabras simples	4. Lectura de palabras simples	Familiar word reading	FWR
Decodificación de palabras sin sentido	Fónica de seudo palabras	50	Precisión o número de pseudo palabras leídas correctamente, velocidad o número de pseudo palabras leídas correctamente por minuto	5. Decodificación de palabras sin sentido	5. Lectura de palabras sin sentido	Nonsense word reading	NWR
Lectura de un pasaje	Fluidez de palabras	64 pero no registrados	Fluidez o número de palabras leídas correctamente por minuto	6. Lectura y comprensión de un pasaje	6. Lectura y comprensión de un pasaje	Connected text fluency	CTF

Aspecto medido	Habilidades	Elementos (ítems)	Formas de cuantificación	Nombre de la sección en Nicaragua	Nombre de la sección en Honduras	Name in English	Abr
Comprensión de lectura de un pasaje	Comprensión de Lectura	5	Porcentaje de respuestas correctas de 5 preguntas	6. Lectura y comprensión de un pasaje	6. Lectura y comprensión de un pasaje	Reading comprehension	RC
Compresión oral de un pasaje	Comprensión oral	5 ó 3	Porcentaje de respuestas correctas de 5 preguntas (caso Honduras) o 3 preguntas (caso Nicaragua)	7. Comprensión oral	7. Comprensión oral	Oral comprehension	OC
Escritura de una oración	Ortografía	7 ó 6	Porcentaje de respuestas correctas de 7 preguntas (caso Honduras) o 6 preguntas (caso Nicaragua)	8. Dictado	8. Dictado	Write dictate text	WDT

Mayores detalles para las formas de calificación de cada sub prueba se encuentran en los respectivos informes de resultados de EGRA para Honduras y Nicaragua (CIASES y RTI International, 2009; CIASES, 2010).

- Comprensión de lectura de un pasaje.
- Comprensión oral de un pasaje.
- Escritura de una oración dictada.

Debemos advertir que los nombres atribuidos a los aspectos medidos pueden ser arbitrarios y no necesariamente corresponden a los nombres que aparecen en los mencionados informe de resultados realizados con los datos de EGRA para estos países, sin embargo estos nombres en lo posible se basan en los nombres en inglés que toman estos aspectos cuando se enfatiza en la tarea asignada. También se esboza para los aspectos medidos a que habilidades corresponden no significando que estos aspectos medidos cubren exhaustivamente las habilidades indicadas.

A continuación describimos estas mediciones.

Ubicación espacial para leer un párrafo: consiste en tres preguntas que permiten conocer si el estudiante se orienta adecuadamente para la lectura de un párrafo; es decir, si demuestra con el dedo dónde comenzar a leer, cómo seguir la lectura de la primera línea y luego continuar con las siguientes. Se registra si la acción (poner el dedo en la primera palabra, indicar con el dedo el movimiento de derecha a izquierda, pasar el dedo de arriba a abajo) es correcta o incorrecta.

En español, la lectura de un párrafo se inicia en la parte superior izquierda de un párrafo y se realiza de izquierda a derecha, y de arriba a abajo cuando acaba la lectura de la primera línea. Otro nombre que puede usarse para lo que se mide en esta sección es "conocimiento de la dirección para la lectura" (reading direction). Este subtest resulta extremamente importante cuando se comparan los resultados de comprensión lectora entre diferentes lenguas o en situaciones de multilingüismo.

Se espera que todos los estudiantes reconozcan la dirección correcta para la lectura o, en todo caso, que las proporciones de casos en que no se hace disminuyan en los grados escolares superiores. Adicionalmente se puede asumir que este conocimiento ocurre cuando las tres condiciones son satisfechas, es decir, cuando se conoce dónde comenzar, cómo seguir en una línea y cómo seguir entre líneas.

Esta sección de EGRA solamente está presente en la versión aplicada a Nicaragua como sección 1 y es llamada allí "empeño y relación a la letra impresa".

Nombramiento de las letras: consiste en la presentación de 100 letras, las cuales deben ser identificadas en un límite de un minuto por los examinados. Existe la opción de interrumpir la aplicación del subtest, si ninguna de las 10 letras en la primera línea es identificada correctamente. Son registrados principalmente el tiempo empleado por el alumno, si es menos de 60 segundos (si el alumno las identifica todas, el tiempo es implícitamente 60 segundos, ya que ese es el tiempo límite), y el número total de letras identificadas correctamente en el tiempo considerado. La primera medida: el número total de letras nombradas correctamente, corresponde a la precisión y la segunda medida: el número total de letras nombradas por minuto, corresponde a la rapidez o automaticidad.

La segunda medida está relacionada con la fluidez con que se reconocen las letras. Los totales de letras correcta e incorrectamente leídas son anotados con posterioridad a la culminación de la aplicación.

El alfabeto español consta de 29 símbolos (27 letras y 2 dígrafos, Ch y Ll). La lectura de las letras del alfabeto debe hacerse usando tanto minúsculas como mayúsculas. Otro nombre que puede usarse para lo que se mide en esta sección es "reconocimiento del nombre de las letras" (letter-name recognition). Este subtest resulta extremamente importante porque mide aspectos de fluidez o automaticidad que favorecen la lectura.

Se espera que todos los estudiantes identifiquen correctamente las letras, pronunciando el nombre correcto de estas en el alfabeto español. Se considera que se encuentran alfabetizados, por estar mínimamente en el 2º grado escolar.

En todo caso, se espera que el número de nombres de letras reconocidas o la velocidad de reconocimiento del nombre de las letras aumente conforme aumenta el grado escolar. Adicionalmente se espera que este reconocimiento ocurra independientemente del uso de letras mayúsculas o minúsculas y del orden en que estas aparecen.

Esta sección de EGRA consta de 50 letras minúsculas y 50 mayúsculas en la aplicación en Nicaragua, y de 42 minúsculas y 58 mayúsculas en el caso de Honduras; corresponde a las secciones 2 y 1 respectivamente, con el nombre "conocimiento de las letras". La aplicación en Nicaragua emplea dos ejemplos preliminares (letras C y T), en tanto en Honduras se emplean 3 (letras F, T, a). También el orden de presentación de las letras difiere entre ambas aplicaciones y, finalmente, en EGRA Honduras la letra w no es presentada.

Identificación del sonido de las letras: consiste en la presentación de 50 letras ; el sonido de cada una debe ser producido en un tiempo límite de un minuto por los examinados. Existe la opción de interrumpir la aplicación del subtest si ninguno de los elementos de la primera línea de 10 letras pudo ser completada correctamente. Se registra el tiempo empleado por el alumno si es menor a 60 segundos, y el total de letras presentadas en el tiempo considerado. En la primera medida, el número de sonidos identificados correctamente corresponde a la precisión, y en la segunda medida, el total de sonidos identificados correctamente por minuto corresponde a la rapidez o automaticidad. Ambos aspectos están relacionados con la habilidad para emparejar letras (grafemas) con sonidos (fonemas), que son parte del sistema fonológico de un idioma. El número total de sonidos identificados correctamente es anotado con posterioridad a la culminación de la aplicación.

El alfabeto español consta de 29 símbolos (27 letras y 2 dígrafos, ch y ll). La pronunciación del sonido de las letras del alfabeto no es diferente si están escritas con minúsculas o con mayúsculas. Otro nombre que puede usarse para lo que se mide en esta sección es "identificación del sonido de las letras" (letter-sound identification). Este subtest resulta extremamente importante, porque mide habilidades que favorecen la lectura.

Se espera que todos los estudiantes produzcan correctamente los sonidos de las letras del alfabeto español, considerando que se espera que estén alfabetizados, por estar mínimamente en el 2º grado escolar. En todo caso, se espera que el número de sonidos de letras identificados o la velocidad para el sonido de las letras recordadas aumente conforme el grado escolar. Adicionalmente, se espera que esto ocurra independientemente del uso de letras mayúsculas o minúsculas y del orden en que estas aparecen.

Esta sección de EGRA solamente está presente en la versión aplicada a Honduras como sección 3, con el nombre "conocimiento del sonido de las letras" y consta de 26 letras minúsculas y 24 mayúsculas.

Identificación del sonido de la letra inicial de una palabra: está formado por 10 palabras: el examinado debe identificar el primer sonido de cada palabra dada por el examinador en un máximo de 15 segundos. Existe la opción de interrumpir la aplicación del subtest si no se reconoce correctamente el sonido de la primera letra en ninguna de las 5 primeras palabras leídas. Se registra si el ítem se completó correctamente, incorrectamente o si el examinado se rehusó a responder. El número total de sonidos iniciales identificados correctamente es anotado con posterioridad a la culminación de la aplicación.

Se espera el reconocimiento del sonido de la letra consonante que inicia una palabra cuando esta es leída. Otro nombre que puede usarse para lo que se mide en esta sección es "reconocimiento del sonido de la letra inicial de una palabra escuchada" (letter-sound recognition by word). Este subtest resulta extremamente importante porque mide la conciencia fonémica que favorece la lectura.

Se espera que la mayoría de los estudiantes identifique correctamente el sonido de la letra inicial de las palabras presentadas tomando en cuenta que la mayor parte de ellas son de uso cotidiano en niños que están mínimamente en el 2º grado escolar. En todo caso, se espera que el número de sonidos de letras reconocidas o la velocidad de reconocimiento del sonido de las letras aumente conforme avanza el grado escolar. Adicionalmente, se espera que este reconocimiento ocurra independientemente del tamaño de la palabra.

Esta sección de EGRA consta de 10 palabras en ambas aplicaciones, con dos palabras diferentes entre ellas, y corresponde a la sección 2 en Honduras con el nombre "sonido inicial", y a la sección 3 en Nicaragua, bajo el nombre "conciencia fonémica y fonológica.

En ambas aplicaciones se cuenta con dos ejemplos, uno de los cuales es diferente en cada caso.

Identificación de palabras que se inician con el mismo sonido: consiste en la lectura de 10 juegos de tres palabras, en los cuales el examinado debe identificar la palabra que presenta diferente sonido de la letra inicial en un máximo de 15 segundos. Existe la opción de interrumpir la aplicación del subtest si no reconoce correctamente ninguno de las palabras con sonidos diferentes en la letra inicial en los 5 primeros juegos leídos. Se

registra si la actividad se hizo correctamente, incorrectamente o si el examinado se rehusó a responder. El total de palabras con sonido diferente en la primera letra reconocidos, tanto correcta como incorrectamente, son anotados con posterioridad a la culminación de la aplicación.

El alfabeto español consta de 22 consonantes y 5 vocales. Se espera el reconocimiento del sonido de la letra consonante diferente que inicia una palabra cuando esta es leída junto a otras dos iguales. Otro nombre que puede usarse para lo que se mide en esta sección es "reconocimiento del sonido de la letra inicial diferente de una palabra de tres palabras leídas" (letter-sound recognition between word). Este subtest resulta extremamente importante porque mide la conciencia fonológica, que favorece la lectura.

Se espera que la mayoría de los estudiantes identifiquen correctamente el sonido de la letra inicial entre las palabras presentadas, tomando en cuenta que la mayor parte de ellas son de uso cotidiano en niños que están mínimamente en el 2º grado escolar. En todo caso, se espera que el número de sonidos de letras reconocidas o la velocidad de reconocimiento del sonido de las letras aumente conforme avanza el grado escolar. Adicionalmente, se espera que este reconocimiento ocurra independientemente del tamaño y la ubicación de la palabra con el sonido de la primera letra diferente.

Esta sección de EGRA solamente está presente en la versión aplicada a Nicaragua con el nombre de "consciencia fonémica y fonológica" y cuenta con dos ejemplos previos.

Lectura de palabras simples: consiste en la presentación de 50 palabras, las cuales deben ser leídas por los examinados en un tiempo límite de un minuto. Existe la opción de interrumpir la aplicación del subtest si ninguna de las palabras en la primera línea de 5 palabras pudo ser completada correctamente. Son registrados, principalmente, el tiempo empleado por el alumno, si es menor a 60 segundos; y el total de palabras leídas en el tiempo considerado. La primera medida: total de palabras simples leídas correctamente, corresponde a la precisión, y la segunda medida: total de palabras simples leídas correctamente por minuto, corresponde a la rapidez o automaticidad. Los totales de palabras correcta e incorrectamente leídas son anotados con posterioridad a la culminación de la aplicación.

Las palabras aparecen en minúsculas y son de 2 a 5 letras. Otro nombre que puede usarse para lo que se mide en esta sección es "lectura de palabras simples" (familiar word reading). Este subtest resulta extremamente importante, porque mide la fonética y el reconocimiento automático de palabras comunes, ambos de los cuales favorecen la lectura.

Se espera que todos los estudiantes lean correctamente las palabras en la lista presentada, considerando que se encuentran alfabetizados por estar mínimamente en el 2º grado escolar. En todo caso, se espera que el número de palabras leídas o la velocidad de lectura de las palabras aumente conforme avanza el grado escolar. Adicionalmente, se espera que la lectura de las palabras ocurra independientemente del tamaño y orden en que estas aparecen.

En ambos países esta sección de EGRA es la número 4, con el nombre de “conocimiento de palabras simples”, y consta de 50 palabras en minúscula.

Decodificación de palabras sin sentido o palabras imaginarias: consiste en la presentación de 50 pseudo palabras que deben ser leídas por el examinado en un tiempo límite de un minuto. Existe la opción de interrumpir la aplicación del subtest si ninguna de las palabras en la primera línea de 5 pseudo palabras pudo ser completada correctamente. Son registrados principalmente el tiempo empleado por el alumno, si es menor a 60 segundos, y el total de pseudo palabras leídas en el tiempo considerado. La primera medida: total de pseudo palabras leídas correctamente, corresponde a la precisión, y la segunda medida: total de pseudo palabras leídas correctamente por minuto, corresponde a la rapidez.

Los totales de pseudo palabras correcta e incorrectamente leídas son anotados con posterioridad a la culminación de la aplicación.

Las pseudo palabras aparecen en minúsculas y son de 4 a 5 letras. Otro nombre que puede usarse para lo que se mide en esta sección es “lectura de palabras sin sentido” (nonsense word reading o pseudoword decoding). Este subtest resulta extremamente importante porque mide la fonética, lo cual favorece la lectura.

Se espera que todos los estudiantes lean correctamente las palabras en la lista presentada, considerando que se encuentran alfabetizados por estar mínimamente en el 2º grado escolar. En todo caso, se espera que el número de pseudo palabras leídas o la velocidad de lectura de estas aumenten con el grado escolar. Adicionalmente, se espera que esta lectura ocurra independientemente del tamaño y orden en que las palabras aparecen.

En ambos países esta sección de EGRA es la número 5, con el nombre “decodificación de palabras sin sentido”, y contiene 2 palabras diferentes y 48 comunes.

Lectura de un pasaje: contiene un texto de 64 palabras que debe ser leído por los examinados en un tiempo límite de un minuto. Existe la opción de interrumpir la aplicación del subtest si ninguna de las palabras pudo ser completada correctamente en la primera línea de 11 palabras (caso Nicaragua) o de 4 palabras (caso Honduras). Son registrados, principalmente, el tiempo empleado por el alumno, si es menor de 60 segundos, y el total de palabras leídas en el tiempo considerado. La primera medida, el total de palabras correctas, corresponde a la precisión, y la segunda, al número de palabras leídas correctamente por minuto corresponde a la rapidez o automaticidad. La segunda medida está relacionada con la fluidez con que se leen palabras en un pasaje. Los totales de palabras correcta e incorrectamente leídas son anotados con posterioridad a la culminación de la aplicación.

Otro nombre que puede usarse para lo que se mide en esta sección es “lectura de palabras en un texto” (connected text fluency o oral reading fluency in connected text) Este subtest resulta extremamente importante porque mide habilidades de fluidez que favorecen la lectura.

Se espera que los estudiantes lean correctamente las palabras del texto presentado, considerando que se encuentran alfabetizados por estar mínimamente en el 2º grado escolar. En todo caso, se espera que el número de palabras leídas o la velocidad de lectura de las palabras en el texto aumente con el grado escolar. Adicionalmente, se espera que esta lectura ocurra independientemente del tamaño y orden en que las palabras aparecen.

En ambos países esta sección de EGRA es la número 6, con el nombre de "lectura y comprensión de un pasaje". En la aplicación de Honduras el texto es presentado en 11 líneas, mientras que en Nicaragua se presenta en 6.

Comprensión de lectura de un pasaje: a partir de un texto presentado, en el subtest de lectura de un pasaje se presentan cinco preguntas relacionadas con el texto. Como en el caso de la lectura del texto, se mantiene la opción de interrumpir la aplicación del subtest si ninguna de las palabras en la primera línea de 11 palabras (caso Nicaragua) o de 4 palabras (caso Honduras) pudo ser completada correctamente. Se registra si la respuesta se hizo correctamente, incorrectamente o si el examinado se rehusó a responder. Los totales de respuestas correctas o incorrectas son anotados posteriormente.

Otro nombre que puede usarse para lo que se mide en esta sección es "comprensión de lectura" (reading comprehension). Este subtest resulta extremamente importante porque mide habilidades de comprensión de la lectura.

Se espera que los estudiantes respondan correctamente las preguntas relacionadas con la lectura presentada, considerando que se encuentran alfabetizados por estar mínimamente en el 2º grado escolar. En todo caso, se espera que el número de respuestas correctas aumente con el grado escolar. Adicionalmente se espera que esta comprensión ocurra independientemente de las preguntas formuladas en relación con la lectura.

Esta sección de EGRA es parte de la sección 6 en ambos países con el nombre de "lectura y comprensión de un pasaje", y presenta una pregunta común y cuatro preguntas diferentes entre ellas.

Comprensión oral de un pasaje: es leído un texto de 51 palabras (caso Honduras) o de 29 palabras (caso Nicaragua) y luego se hacen preguntas relacionadas con este, 5 en el caso de Honduras y 3 en el caso de Nicaragua. No existe opción de interrumpir el subtest. Se registra si la respuesta fue correcta, incorrecta, o si el examinado se rehusó a responder. Los totales de respuestas correctas o incorrectas son anotados posteriormente.

Otro nombre que puede usarse para lo que se mide en esta sección es "comprensión oral" (oral comprehension). Este subtest resulta extremamente importante porque mide habilidades que son netamente pre-lectura, pero que favorecen la lectura.

Se espera que los estudiantes respondan correctamente las preguntas relacionadas con la lectura presentada, considerando que se encuentran alfabetizados por estar mínimamente en el 2º grado escolar. En todo caso, se espera que el número de respuestas correctas aumente con el grado escolar. Adicionalmente, se espera que esta comprensión ocurra independientemente de las preguntas formuladas en relación con la lectura.

Esta sección de EGRA es parte de la sección 7 en ambos países, con el nombre de "comprensión oral", y presenta una pregunta común y cuatro preguntas diferentes entre ellas.

Escritura de una oración dictada: es leída una oración de 14 palabras (caso Honduras) y de 8 palabras (caso Nicaragua) para que el examinado escriba. No existe opción de interrumpir el subtest. Se registra si la respuesta se hizo nada correcta, algo correcta y correcta, evaluando la escritura de algunas palabras (4 en Honduras y 3 en Nicaragua), el uso de espacio y dirección de texto, el uso de mayúsculas y puntuación correcta. Los totales de respuestas correctas o incorrectas son anotados posteriormente.

Otro nombre que puede usarse para lo que se mide en esta sección es "escritura de una oración dictada" (ability to write dictated text).

Se espera que los estudiantes escriban correctamente la oración dictada, considerando que se encuentran alfabetizados por estar mínimamente en el 2º grado escolar. En todo caso, se espera que el puntaje alcanzado aumente conforme el grado escolar. Adicionalmente, se espera que este manejo ocurra independiente de las oraciones dictadas.

Esta sección de EGRA es parte de la sección 8 en ambos países, con el nombre de "dictado", y es diferente, en oración y evaluación de lo escrito en ambos países.

Formas de calificación

Adicionalmente el cuadro 1 más arriba indica las formas de cuantificación posibles en cada aspecto medido. Siendo el caso que algunos aspectos medidos pueden ser cuantificados en más de una manera. En el cuadro 1 se describen los aspectos medidos en EGRA, la definición operacional de estos aspectos, el número de elementos y las formas de cuantificación.

Usos y reporte de resultados de EGRA

Los resultados completos de las aplicaciones de EGRA en Honduras y Nicaragua se encuentran en reportes para cada país, disponibles en www.eddataglobal.org. Por favor, consultar los reportes para conocer los resultados completos. Para revisar extensiones de uso de EGRA puede revisarse Roskos, Strickland, Haase y Malik (2009) y Gove y Cvelich (2010).

Para una revisión acerca de la importancia de la medicion de habilidades de lectura en los niveles iniciales de la educación primaria y su impacto para la comparabilidad entre países puede revisarse.

1.2.2 La evaluación psicométrica de un instrumento

Todo proceso de desarrollo y construcción de pruebas pasa por diversas etapas. Estas pueden incluir el planeamiento de la prueba, la selección de áreas que se incluirán en la prueba, la proposición de un conjunto de preguntas que cubren las áreas elegidas, la

administración de una prueba piloto para el ensayo de las preguntas seleccionadas, el proceso de análisis de las preguntas (lo que lleva a la selección de las mejores preguntas) y una administración final sobre la base de una muestra que servirá para la versión final de la prueba.

Sin embargo, como ocurre en diferentes sistemas de evaluación públicos y privados, regionales, nacionales e internacionales, una etapa previa del proceso de estimación de las habilidades definitivas de los estudiantes que respondieron las pruebas es evaluar las propiedades o características psicométricas de la versión final de las mismas. Esta evaluación es importante porque resulta conveniente establecer las propiedades finales de las pruebas y preguntas en el marco de la población objetivo considerada.

Como mencionan Bazán y Millones (2002a), una de las primeras preguntas que debe resolverse en el análisis de las pruebas se refiere a lo que se espera del conjunto de las pruebas. Es decir, qué se pretende hacer con los resultados, qué tipo de conclusiones podremos generar de ellas, lo cual constituye un aspecto clave para determinar una de las propiedades que debe cumplir la prueba: su validez.

Por ejemplo, las pruebas de rendimiento pueden ser diseñadas para evaluar el cambio relativo experimentado en los puntajes promedio de los estudiantes de unas escuelas entre el principio y el final de un determinado año escolar. En este caso, el modelo adoptado, el modelo de normas, permite establecer una norma o escala de referencia (prueba de entrada), con el objetivo de obtener una calificación para poder compararla con una evaluación posterior (prueba de salida). De esta manera se puede determinar cuánto han avanzado los estudiantes respecto a ellos mismos o a su grupo de referencia. Esto es posible porque la comparación considera al estudiante respecto a sí mismo y no respecto a un estándar o criterio de excelencia, como ocurre con la interpretación de Criterios[4].

La interpretación de normas de una prueba busca estimar las distancias relativas entre grupos de interés o sub poblaciones. Estas comparaciones pueden definirse de varias formas. Así, para algunos usos es interesante referirse a una norma nacional o regional. En otros casos se puede hacer referencia a sub distribuciones de grupos más específicos, como los que podrían ser los de la gestión pública o privada, etc. Las interpretaciones basadas en tales comparaciones suelen tener su referente en normas. Estadísticas adicionales, como percentiles o cuartiles de la distribución relevante, son útiles para especificar mejor las comparaciones entre estos grupos de interés.

Por el contrario, en la interpretación de criterios no se hace referencia directa al desempeño de otros examinados. El interés se centra en determinar la probabilidad de éxito (individual) respecto a algún dominio de preguntas. Este tipo de interpretación también toma una variedad de formas. Por ejemplo, es posible referirse a la probabilidad de éxito o de respuesta correcta en un subconjunto de preguntas de la prueba (un dominio

[4] Esto no quiere decir que un modelo de criterios no puede ser adoptado para evaluar cambio. Por el contrario, el modelo de criterios puede ser conveniente cuando se desea evaluar los cambios de aprendizaje experimentados por la población evaluada.

de la prueba) o para un dominio más amplio de preguntas. Así, si se emplea este tipo de interpretación para evaluar cambios entre los estudiantes en dos ocasiones, el modelo de criterios ofrece una mejor manera de evaluar los cambios en el aprendizaje considerando un estándar.

De otro lado, en lo que se refiere a habilidades de lectura, generalmente las pruebas buscan determinar el nivel de desempeño del estudiante en relación con un determinado estándar para su nivel educativo. Ese estándar puede ser determinado de diversas maneras para reflejar aquello que se espera en el correspondiente nivel educativo considerado. En este caso, las pruebas referidas a criterios resultan ser más apropiadas. De esta manera, el estudiante queda clasificado en un determinado estado de maestría o categoría en relación con aquello que está siendo evaluado (Swaminathan, Hambleton y Algina, 2005).

Un segundo aspecto para la definición de las propiedades psicométricas de las pruebas está relacionado con el marco referencial teórico (modelo de medición) que se adopta en el análisis de las pruebas en sus diversas etapas. El modelo de medición debe ayudar a comprender y a evaluar los puntajes que vienen de las respuestas a los ítems y de aquí hacia el constructo, y este también debe guiar el uso de los resultados en aplicaciones prácticas. Simplemente, el modelo debe traducir puntajes de respuesta a localizaciones en el mapa de constructo (Wilson, 2005).

Como se sabe, los modelos de medición más usados son los de la Teoría Clásica de los Tests y los modelos de la familia de la Teoría de Respuestas a Ítems o TRI, el modelo de Rasch es un caso especial.

La Teoría Clásica de los Tests (Lord y Novick, 1968) es un enfoque según el cual el resultado de la medición de una variable depende de la prueba utilizada y de los sujetos evaluados, y normalmente se ofrece el resultado de un ítem como el porcentaje respondido correctamente. El total de los resultados se puede presentar como el total de ítems respondidos correctamente, lo cual en cierta medida facilita su interpretación. Sin embargo, el énfasis que pone esta teoría en las pruebas utilizadas ha sido causa de críticas, pues en dicha estrategia una variable es inseparable del instrumento utilizado para medirla y ello constituye una seria limitación, ya que inevitablemente se acabaría definiendo operativamente la variable por el instrumento con que se mide. En otras palabras, uno de los problemas con el análisis clásico es que no hay forma de definir la dificultad de un ítem que no se auto-refiera al porcentaje de alumnos que lo responden correctamente: es una definición hasta cierto punto circular.

Por otro lado, la Teoría de la Respuesta al Ítem (Baker y Kim, 2004) es un importante modelo que enfatiza el análisis de las preguntas, estableciendo un conjunto de supuestos acerca del comportamiento de los sujetos al responderlas y determinando un modelo probabilístico para las respuestas correctas e incorrectas a partir de la consideración de algunos parámetros relativos a las preguntas y una variable latente subyacente para la habilidad de los evaluados. En el caso del modelo de Rasch (Bond y Fox, 2001) el único

parámetro asociado a las preguntas es la dificultad, la cual se encuentra en la misma escala que las habilidades.

En general, las pruebas, sean para su interpretación en normas o criterios, deben cumplir con un conjunto de requisitos para establecer la validez y confiabilidad de las mismas, aspectos que abordamos a continuación. Esto determina que es necesario establecer diferentes criterios de evaluación psicométrica que consideren una decisión acerca del modelo que se emplea.

Los *Standards for educational and psychological testing* (AERA, APA, NCME, 1999) y Joint Committee on Standards for Educational Evaluation (2003), entre otros, han descrito métodos que pueden ser usados para obtener evidencia para la confiabilidad y validez de pruebas, aspectos que abordamos a continuación.

Validez de pruebas

1. Definición

La validez se refiere al grado por el cual la evidencia y la teoría respaldan las interpretaciones de los puntajes de las pruebas. La validez es el aspecto más importante en el desarrollo y evaluación de estas.

El proceso de validación involucra la acumulación de evidencia para proveer una base científica para las propuestas de interpretación de los puntajes de la prueba. Así, se deben evaluar los usos propuestos y no la prueba en sí misma.

Cuando los puntajes de las pruebas se usan o interpretan de maneras diferentes, cada una de estas interpretaciones debe ser validada. Por ejemplo, una prueba de rendimiento en español puede ser usada con uno de los siguientes fines: ubicar a un estudiante en un programa de instrucción apropiado, aprobar al estudiante en el curso correspondiente, o seleccionar al estudiante para un concurso inter-escolar. Cada uno de esos usos implica una determinada interpretación de los puntajes de la prueba; por ejemplo, si el estudiante puede beneficiarse de una particular intervención instruccional, si el estudiante tiene dominio en un currículo especifico o si el estudiante está apto para participar en el concurso inter escolar. Cada uno de esos potenciales usos implica tareas específicas, formas diferentes de interpretación de los puntajes y desarrollo y evaluación diferentes.

Es importante señalar que las consideraciones de validez son distintas si se trata de emitir un juicio sobre individuos o sobre la base de grupos grandes; o sea, si se trata de tomar decisiones sobre el destino de un alumno específico, o si se trata, por el contrario, de la toma de decisiones sobre políticas educacionales, ajuste curricular, etc.

2. Evidencias a considerar

La decisión acerca de qué tipo de evidencias son importantes para validación varían en cada caso. Sin embargo, es natural que algunos tipos de evidencia puedan ser especialmente críticos en un determinado caso, mientras que otros pueden ser menos útiles.

Por ejemplo, si la prueba de rendimiento en español se usa como pre-requisito en un curso avanzado, algunas evidencias pueden ser las siguientes:

- El alumno posee las habilidades que son pre-requisito para el curso avanzado.
- El dominio de contenido de la prueba es consistente con esas habilidades.
- Los puntajes de la prueba pueden ser generalizados a través de un conjunto relevante de ítems.
- Los puntajes de las pruebas no son influenciados por variables externas, tales como la habilidad para escribir.
- El éxito en el curso avanzado puede ser válidamente evaluado.
- Los examinados con puntajes altos en la prueba son más exitosos en el curso avanzado que los examinados con puntajes bajos.

3. Validación de un constructo

La identificación de las proposiciones que surgen de una determinada interpretación de la prueba se pueden facilitar considerando una hipótesis rival que podría cambiar la interpretación propuesta. Es también útil considerar las perspectivas de diferentes partes interesadas, cuando existen experiencias con pruebas, contextos y consecuencias del uso propuesto de pruebas similares.

La sub representación del constructo se refiere al grado por el cual una prueba falla en capturar aspectos importantes del constructo. Por ejemplo, la prueba no contiene el suficiente tipo de contenidos, procesos psicológicos o formatos para medir el rendimiento en español.

La **varianza irrelevante del constructo** se refiere al grado por el cual un puntaje de una prueba es afectado por procesos extraños al constructo intencional. Por ejemplo, la prueba es afectada por componentes emocionales, familiaridad con el tipo de preguntas, etc.

La validación implica poner gran atención sobre las posibles distorsiones en el significado alrededor de una representación inadecuada del constructo, así como en los aspectos de medición tales como el formato, las condiciones de administración o el nivel de lenguaje, todos ellos aspectos que pueden limitar o modificar la interpretación de los puntajes de la prueba.

El proceso de validación puede llevar a revisiones de la prueba, el marco conceptual de la prueba o en ambos. Sin embargo, la prueba revisada debe necesariamente ser validada.

4. Responsabilidad de la validación

La validación es una responsabilidad conjunta del desarrollador de la prueba y del usuario de la prueba. El desarrollador es responsable de dar evidencia relevante y una racionalidad que respalde el uso pretendido de la prueba. El usuario es el responsable final de evaluar la evidencia con respecto al particular propósito de uso de la prueba.

Cuando el uso de una prueba difiere de lo respaldado por el desarrollador, el usuario es especialmente responsable de la validación.

Adicionalmente, importantes contribuciones a la evidencia de validez son dadas por investigadores, que relacionan los puntajes de la prueba con otras variables.

Fuentes de evidencia de validez: Hay diferentes tipos de evidencia de validez, que se pueden clasificar en:

- Evidencias basadas en el contenido de la prueba.
- Evidencias basadas en el proceso de respuesta.
- Evidencias basadas en la estructura interna.
- Evidencias basadas en las relaciones con otras variables.
- Evidencias basadas en las consecuencias del uso de la prueba.

5. Evidencias basadas en el contenido de la prueba

Se obtienen a partir de un análisis de la relación entre el contenido de la prueba y el constructo a ser medido. El contenido se refiere a temas, redacciones, formatos de ítems, tareas o cuestiones en las pruebas, así como a las guías para los procedimientos con respecto a administración y calificación.

La evidencia puede incluir un análisis lógico y empírico de la adecuación, con la cual el contenido de la prueba representa el dominio de contenido, y de la relevancia del dominio de contenido para la interpretación de los puntajes de la prueba. También puede incluir juicio de expertos de la relación entre las partes de la prueba y el constructo.

Evidencias basadas en los procesos de respuesta: Análisis teóricos y empíricos de los procesos de respuesta proporcionan evidencias referidas al ajuste entre el constructo y la naturaleza del desempeño de respuesta que usan los examinados. Por ejemplo, si evaluamos rendimiento en español, los procesos implicados en la respuesta de los ítems deben incluir habilidades en este dominio y no aspectos de memoria.

Este tipo de evidencia generalmente proviene de análisis individuales de respuestas, o de cuestiones referidas a las estrategias de resolución de las preguntas, incluyendo tiempos de respuesta. También se puede obtener evidencia analizando la relación entre partes de la prueba y otras variables que ayudan a reconsiderar formatos.

Estudios de procesos de respuesta que involucran a examinados de diferentes subgrupos, pueden ayudar a determinar la influencia de otras variables en el desempeño durante la prueba.

Las evidencias también pueden incluir juicios de evaluadores acerca de los procesos de respuesta de los examinados.

Evidencias basadas en la estructura interna de la prueba: Estas evidencias pueden indicar el grado en el que la relación entre los ítems de la prueba y los componentes de la prueba conforman el constructo en el cual se basan los puntajes de la prueba propuesta.

El marco conceptual de una prueba puede implicar una simple dimensión o varios componentes que se esperan homogéneos pero distintos entre sí. El tipo de análisis y su interpretación depende de cómo se use la prueba.

Evidencias basadas en relaciones con otras variables: El análisis de los puntajes de la prueba con otras variables externas proporciona una fuente importante de validez. Variables externas pueden ser: medidas del mismo criterio que la prueba espera predecir, otras pruebas que hipotéticamente miden el mismo constructo, pruebas que miden constructos relacionados o diferentes.

Las evidencias que se encuentran abordan cuestiones acerca del grado en el cual esas relaciones son consistentes con otros constructos subyacentes a las interpretaciones de la prueba. Estas evidencias se pueden basar en estudios correlacionales o experimentales.

Evidencia convergente y discriminante: la evidencia convergente es determinada por las relaciones con constructos similares. La evidencia discriminante es determinada por relaciones con constructos diferentes. Por ejemplo, la prueba de rendimiento en español debe estar relacionada con pruebas de razonamiento en español y de comprensión de lectura (evidencia convergente) y menos relacionada con pruebas de razonamiento matemático (evidencia discriminante).

Relaciones de la prueba con un criterio: la cuestión fundamental es: ¿con qué grado de precisión los puntajes de la prueba predicen el desempeño en el criterio?

El grado de precisión que se considera necesario depende del propósito del uso de la prueba. Sin embargo, la variable criterio es la medida de algún atributo o respuesta que es de interés primario para los usuarios de la prueba.

La elección del criterio y los procedimientos de medición usados para obtener puntajes de criterio son de importancia central. Históricamente hay dos diseños: predictivo y concurrente.

Un *estudio predictivo* indica cuán precisamente los datos de la prueba pueden predecir puntajes de criterio que son obtenidos un tiempo después. Un *estudio concurrente* obtiene información del predictor y del criterio al mismo tiempo. Por ejemplo, un estudio predictivo consistiría en hacer un estudio de seguimiento de los ingresantes en el curso Matemática I, considerando los resultados de la prueba de razonamiento matemático tomada en el ingreso.

Un estudio *concurrente* consistiría en analizar el resultado del ingreso (puntaje total) con el resultado en la prueba de razonamiento matemático tomada en el ingreso.

Generalización de la validez: es el grado en el cual la evidencia de validez basada en relaciones de la prueba con un criterio puede ser generalizada a una nueva situación sin hacer estudios adicionales de validez en esta.

Resúmenes estadísticos de estudios de validación en situaciones similares pueden ser útiles en estimar relaciones de la prueba con el criterio en una nueva situación.

Cuando existe evidencia abundante se pueden realizar estudios meta analíticos. Cuando no, hay que realizar procedimientos cuidadosos. Los estudios pueden variar de acuerdo con:

1. diferencias en la manera como se mide el constructo de predicción,
2. el tipo de trabajo o currículo involucrado,
3. el tipo de criterio de medición usado,
4. el tipo de usuarios de la prueba,
5. el período en el cual los estudios fueron realizados.

En cada estudio particular estos aspectos varían y el objetivo es determinar empíricamente la extensión por la cual estas variaciones en los aspectos afectan las correlaciones entre la prueba y el criterio.

Evidencias basadas en las consecuencias de la prueba: En años recientes se ha prestado atención a la incorporación de las consecuencias pensadas y no pensadas de la prueba en el aspecto de validez. Sin embargo, hay que distinguir entre evidencia relevante para la validez y evidencia que puede informar de decisiones acerca la política social.

Aunque la información acerca de las consecuencias de la prueba pueda influir en las decisiones de su uso, tales consecuencias no deben disminuir la validez de las interpretaciones planificadas. Las pruebas se administran, comúnmente, en la expectativa de obtener algún beneficio con sus puntajes. Así, un propósito fundamental de la validación es indicar cómo serán obtenidos estos beneficios. Por ejemplo, en la prueba de rendimiento en español podría haber evidencia acerca de cómo este resultado podría ayudar a promover a los alumnos de grado escolar.

6. Integrando la evidencia de validez

Un argumento de validez legítimo debe integrar varias fuentes de evidencia en una cantidad coherente con el grado en el cual la evidencia existente y la teoría respaldan las interpretaciones pretendidas de los puntajes de la prueba, y abarcar evidencia recogida de nuevos estudios y aquella disponible de reportes de investigación recientes.

El argumento de validez puede indicar la necesidad de refinar la definición del constructo, puede sugerir revisiones en la prueba u otros aspectos del proceso de pruebas y puede indicar áreas que necesitan estudios futuros.

Finalmente, la validez de una interpretación intencional de los puntajes de las pruebas cuenta con toda la evidencia relevante disponible para la calidad técnica del sistema. Esto incluye evidencia de aspectos relativos a la construcción de la prueba, como:

- confiabilidad adecuada
- administración de la prueba apropiada,
- calificación de la prueba apropiada,

- escalamiento de puntajes precisos,
- equivalencias,
- protocolos estandarizados,
- atención cuidadosa respecto a la honestidad de las respuestas de los examinados.

Para asegurar estos aspectos, AERA, APA, NCME (1999) listan 24 estándares de validez que pueden ser explorados en el caso de EGRA.

Confiabilidad de pruebas

Una prueba, definida de manera amplia, es un conjunto de tareas o una escala, diseñadas para describir o hacer explícitas conductas de examinados en un dominio específico, o un sistema para recolectar muestras de trabajos individuales en un área particular. Acoplado a este dispositivo hay un procedimiento de calificación que hace posible que el examinador pueda cuantificar, evaluar, e interpretar las muestras de conducta o trabajo.

La confiabilidad se refiere a la consistencia de tales medidas cuando los procedimientos de *testing* son repetidos en poblaciones de individuos o grupos admitiendo la presencia de un componente de error.

Decir que un puntaje implica un componente de error significa que existe un hipotético valor libre de error que caracteriza a un examinado al momento del *testing*. Por ejemplo, en la Teoría Clásica de los Test este valor es el *puntaje verdadero* (puntaje promedio hipotético resultante de muchas repeticiones de la prueba o formas alternativas del instrumento), pero en la TRI, el valor es referido como parámetro de habilidad o rasgo.

La diferencia hipotética entre el puntaje observado del examinado en cualquier medición particular y el puntaje verdadero o universal es llamada “error de medición”.

Nuevamente, para asegurar estos aspectos, AERA, APA, NCME (1999) listan 20 estándares acerca de confiabilidad y error de medición que pueden ser explorados en el caso de EGRA.

De acuerdo con AERA, APA, NCME (1999), la confiabilidad de una prueba mide el grado en que una prueba es consistente en los puntajes que de ella se obtienen. Idealmente se determina tomando dos o más veces la misma prueba a un examinado y revisando si los puntajes obtenidos son consistentes (idénticos o similares). En la práctica, la consistencia se determina de formas alternativas, una de las cuales se basa en la consistencia interna de la prueba; por ejemplo, cuán consistentemente mide la mitad de una prueba respecto a su otra mitad. Este criterio de consistencia interna de la prueba puede ser calculado por el coeficiente “Alfa” de Cronbach y es reportado comúnmente en diversas evaluaciones de rendimiento, como puede verse en Programme for International Student Assessment (PISA, 2005).

El coeficiente Alfa de Cronbach es un índice que da un valor o cota inferior a la verdadera confiabilidad, pero no es el único. Recientemente ha recibido algunas críticas

dadas por Sijtsma (2009) y por Revelle y Zinbarg (2009), en el sentido de que como medida de confiabilidad y medida de consistencia interna tiene importantes problemas; a) Alfa de Cronbach nunca puede llegar al valor verdadero de confiabilidad, b) en la práctica se usa más como medida de consistencia interna que como medida de confiabilidad, aunque en realidad no está relacionada a la estructura interna de un test.

Zinbarg, Revelle, Yovel y Li (2005) han encontrado que el Alfa de Cronbach puede variar, dependiendo si la escala es unidimensional (un único factor común o un factor general) o multidimensional (muchos factores en común) o si las cargas factoriales en el factor general son todas iguales, lo que torna este índice cuestionable en tales situaciones.

Adicionalmente puede establecerse que: 1) en la práctica el Alfa de Cronbach proporciona valores que están fuera del rango de valores de confiabilidad derivados de una sola administración, 2) Alfa es más usado por estar disponible en softwares comerciales y por ser reconocido en la comunidad académica, en detrimento de otros índices propuestos, 3) Alfa no es el mejor índice para establecer un valor inferior a la verdadera confiabilidad; según Sijtsma (2009), el índice glb es mejor y, según Revelle y Zinbarg (2009), el índice Omega_t es mejor, 4) Alfa no es una medida de consistencia interna ni de unidimensionalidad.

Revelle y Zinbarg (2009) indican que, antes que enfatizar en un índice que proporciona una cota inferior como estimado de la confiabilidad de un test, se debería establecer el porcentaje de un test que mide un constructo; por ejemplo, considerando el índice Omega_h, que mide el factor general de saturación de un test. Así, ellos recomiendan el índice Omega_h, ya que proporciona evidencia de cuánto un test mide un único factor común y esto debe ser obtenido con un análisis factorial jerárquico, considerando una transformación de Schmid-Leiman.

Finalmente, Revelle y Zinbarg (2009) indican que existen cuatro propiedades que un test puede poseer:

1. Unidimensionalidad.- En el sentido de que una única variable latente está siendo medida; tal situación resulta ideal.
2. Homogeneidad o presencia de un factor general.- Los ítems son homogéneos en el sentido de que todos comparten un atributo o variable latente; es decir, si todas las facetas en un dominio están relacionadas al menos con respecto a algún punto, entonces existe una única variable latente que es común a todos los ítems en el dominio. En los casos en que la unidimensionalidad no es realista, se puede probar la homogeneidad, por ejemplo, por medio del análisis factorial confirmatorio de un solo factor general.
3. Factor general de saturación o proporción de varianza del test debida a un factor general.- Es posible que un test sea unidimensional o contenga un factor general, pero que muestre una débil saturación de un factor común lo cual no permita identificar precisamente ni la unidimensionalidad ni el factor general. De esta manera, la proporción de varianza del test debida a un factor general proporciona

importante información acerca del grado en el cual los puntajes totales se generalizan como una variable latente común para todos los ítemes. Esto es medido con el índice Omega_h.

4. Consistencia interna o proporción de varianza del test debida a todos los factores comunes.- Existen algunos contextos, como la predicción aplicada, en los que la preocupación principal consiste en obtener una cota inferior al punto en el cual el puntaje total de un test puede correlacionarse con otras medidas, antes que la identificación teórica de los constructos responsables de tal correlación. Esta proporción da una mayor generalización para identificar el dominio desde el cual los ítems en un test se considera que son una muestra representativa e indica cuáles de estos items pueden representar más de una variable latente. Esto es medido por el índice Omega_t (que es igual a Alfa de Cronbach cuando el test es unidimensional).

Zinbarg et al. (2005) han demostrado que Alfa, Omega_t y Omega_h son equivalentes solo en el caso de unidimensionalidad e igualdad de cargas en el factor general. En otras situaciones, no lo son. Por ejemplo, cuando existe multidimensionalidad y las cargas factoriales en el factor general son desiguales, se espera Alfa < Omega_t, Omega_h < Omega_t. Cuando existe multidimensionalidad y cargas factoriales iguales en el factor general, entonces se espera Omega_h< Alfa < Omega_t. Finalmente, cuando existe unidimensionalidad y cargas factoriales desiguales: Alfa < Omega_h = Omega_t)

La confiabilidad en pruebas relacionadas con un criterio: en pruebas relacionadas con un criterio se producen evaluaciones de la respuesta dada por el estudiante, como ocurre en EGRA. Cuando se toman en cuenta este tipo de juicios, que pueden ser subjetivos, en el puntaje, AERA, APA y NCME (1999) indican que se debe proporcionar evidencia tanto de la consistencia interna en la calificación como de la consistencia dentro de los examinados bajo mediciones repetidas. Se debe hacer una clara distinción entre confiabilidad basada en datos de a) paneles independientes de jueces que califican el mismo desempeño o producto, b) un panel simple que califica desempeños sucesivos o nuevos productos y c) paneles independientes que califican desempeños sucesivos o nuevos productos. En el caso de EGRA, la aplicación implica un proceso activo por parte de los examinadores, que deben cuantificar las respuestas incorrectas de los examinados. Por ello es necesario contar con índices de acuerdo entre examinadores y evidencias de estabilidad de los resultados como es estudiada por ejemplo en Bailey y Bricker (1986).

La confiabilidad en pruebas con límite de tiempo: existe literatura psicométrica que refieren el efecto del límite de tiempo en la confiabilidad de pruebas. Por ejemplo Crocker y Algina (1986) dijeron:

> *When a test has a rigid time limit such that some examinees finish but others do not, an examinee's working rate will systematically influence his or her performance on all forms of the test. . . . On power types of tests, time limits should be long enough to allow all, or nearly all, examinees to finish. Otherwise,*

> *the reliability estimate may be artificially inflated because of consistencies in performance caused by the test's time limit.* (p. 145, citado en Atali, 2005)

De esta manera, se considera que las pruebas con límite de tiempo presentan valores de confiabilidad sobreestimados de los verdaderos. Sin embargo, al contrario de esta creencia común, los estimados de confiabilidad de pruebas de múltiple elección con número de aciertos no están inflados por efecto de *speededness*, como recientemente se ha demostrado (Atali, 2005). Esto se explica porque los examinados responden al azar las preguntas que están fuera del tiempo, de manera que las respuestas a esas preguntas generalmente son menos consistentes con las respuestas a las otras preguntas, con lo cual la confiabilidad del test puede decrecer.

Diferenciabilidad y unidimensionalidad de las pruebas

El análisis de unidimensionalidad de las preguntas que componen las pruebas:

El propósito principal de las pruebas es estimar el rendimiento de los estudiantes sometidos a ellas. En el modelo de Rasch, así como en los modelos de Teoría de Respuesta al Ítem, suponemos que el rendimiento es una variable latente subyacente al conjunto de respuestas dados a las preguntas, que el estudiante responde correcta o incorrectamente. Un nombre genérico adoptado en estas situaciones es el de "habilidad" o "desempeño". Así, con las pruebas consideradas suponemos que existe una habilidad o desempeño, los cuales serán estimados a través del modelo.

En el esquema moderno del concepto de validez se incluye la evidencia de unicidad, es decir, la propiedad de una prueba de medir únicamente un constructo (unicidad de la prueba medible) o desempeño, esto es, establecer si el conjunto de preguntas dentro de una prueba mide una sola cosa—es decir evaluar la unidimensionalidad. No siempre es posible determinar que este supuesto se cumpla cabalmente pero, como Kolen y Brennan (2004) señalan, "aunque en TRI la unidimensionalidad y la independencia local son supuestos que no ocurren estrictamente en la práctica, ellos deben ocurrir lo más cerradamente posible para que la TRI sea usada ventajosamente en muchas situaciones prácticas" (pag 157, traducción propia).

En la literatura psicométrica reciente sobre unidimensionalidad se han desarrollado pruebas específicas para evaluar este aspecto, como por ejemplo la citada por Christensen (2005)[5]. Una prueba estadística específica que se puede considerar para esta evaluación es el test de Martin-Löf (Christensen, Bjorner, Kreiner y Petersen, 2002) el cual está disponible a través de una macro del programa SAS en Christensen y Bjorner (2003). No obstante, es importante considerar más de un criterio para el análisis de unidimensionalidad. Así, uno complementario bastante usado consiste en realizar un análisis factorial a partir de la matriz de correlaciones tetracóricas (Knol & Berger, 1991), considerando que para variables binarias no es posible realizar un análisis factorial usual.

[5] Una interesante discusión acerca de la unidimensionalidad y de las maneras de evaluarla puede verse también en Burga (2005) y en Abedi (1997).

De acuerdo con los resultados, puede considerarse que una prueba será unidimensional si el primer factor explica por lo menos el 40% de la varianza (Carmines y Zeller, 1979). Otro criterio considerado consiste en tomar los valores propios superiores a 1 y realizar un análisis de *scree plot*, identificando las diferencias entre el primer factor y los otros factores con valores propios superiores a 1. Sin embargo, un conjunto de preguntas podría tener múltiples valores propios superiores a 1 y aún ser lo suficientemente unidimensional como para ser analizado con un modelo de teoría de respuesta al ítem (Orlando, Sherbourne y Thissen, 2000). Así, dichos autores consideran que si el número de ítems con cargas factoriales superior a 0,35 es bastante alto, esto puede ser considerado como una evidencia aceptable de unidimensionalidad.

Dado que la unidimensionalidad en el modelo de Rasch dicotómico es el supuesto más importante, en general, para tomar una decisión acerca de la violación de este supuesto es importante considerar no uno, sino varios criterios.

1.3 Metodología del estudio

A continuación describimos la metodología de estudio, indicando las muestras que serán analizadas, los instrumentos considerados y los procedimientos específicos de la evaluación psicométrica que se realizará.

1.3.1 Definición de las muestras de estudio

En el cuadro 2 se describen, por grado y países, las muestras que serán consideradas en este estudio. Se incluye también los casos donde se cuenta con respuestas para EGRA y para pruebas de español en ambos países. En algunos casos aún más específicos, se pudo contar con las respuestas por ítem de dichas pruebas.

Considerando las distinciones realizadas acerca de la versión de EGRA para Nicaragua y para Honduras, ambas pruebas serán analizadas por separado.

Cuadro 2. Muestras para la evaluación psicométrica de EGRA

País	Grado	EGRA	EGRA y prueba de español a nivel de puntajes	EGRA y prueba de español a nivel de ítems
Nicaragua	2º	2164	NA	NA
	3º	2218	NA	NA
	4º	2267	374	374
	Total	6649	374	374

País	Grado	EGRA	EGRA y prueba de español a nivel de puntajes	EGRA y prueba de español a nivel de ítems
Honduras	2º	615	262	243
	3º	597	213	159
	4º	526	265	12
	Total	**1738**	**738**	**414**

Las celdas con NA indican las situaciones donde no se cuenta con una muestra de examinados.

Las muestras que se considerarán para análisis son:

Muestra 1: estudiantes que dieron la versión de EGRA Nicaragua (n=6649).

Muestra 2: estudiantes que dieron la versión de EGRA Honduras (n=1738)

Muestra 3: estudiantes que dieron la versión de EGRA Nicaragua y la prueba de español (n=374)

Muestra 4: estudiantes que dieron la versión de EGRA Honduras y la prueba de español (n=716)

1.3.2 Intrumentos considerados

Para ver los instrumentos considerados en el análisis de este reporte, por favor, consultar www.eddataglobal.org y específicamente CIASES y RTI International (2009) y CIASES (2010).

1.3.3 Procedimientos para la evaluación psicométrica de EGRA

Para este trabajo se proponen algunos criterios para la evaluación psicométrica de EGRA en la línea de trabajo de Bazán y Millones (2002a, 2002b) y el marco conceptual presentado en la sección 1.2.2. Para establecer la validez de las pruebas proponemos los siguientes análisis:

1. Análisis de confiabilidad de los aspectos medidos en EGRA en las versiones de Nicaragua y Honduras, considerando las muestras 1 y 2.
2. Análisis de la estructura correlacional de EGRA en las versiones de Nicaragua y Honduras, considerando las muestras 1 y 2.
3. Análisis de la estructura (consistencia interna y homogeneidad) de EGRA en las versiones de Nicaragua y Honduras, considerando las muestras 1 y 2.
4. Análisis de validez concurrente de EGRA en las versiones de Nicaragua y Honduras, considerando las muestras 3 y 4.

Capítulo 2. Resultados

2.1 Análisis de la confiabilidad de los aspectos medidos en EGRA

El cuadro 3 siguiente muestra los valores de Alfa de Cronbach de los diferentes aspectos medidos en las versiones de EGRA de Nicaragua y Honduras. Nótese que en algunos casos los aspectos medidos son comunes en ambas versiones, pero en otros casos se midieron únicamente en una de las versiones. Los valores de Alfa de Cronbach también se presentan separados por grados.

Debemos indicar que no es posible reportar una medida de confiabilidad específica para lectura de un pasaje en EGRA de ambos países; por ejemplo, utilizando el coeficiente Alfa de Cronbach, debido a que únicamente se registra el total de palabras leídas, correctas e incorrectas, y no se cuenta con un registro por cada palabra. No obstante, este subtest es parte importante de una medida general basada en EGRA, como será reportado en la sección 2.4.

De un modo general, un test es clasificado considerando que tiene confiabilidad apropiada cuando el valor de Alfa es por lo menos de 0.70 (Nunnally, 1978). También Murphy y Davidshofer (1988, p. 89) presentan una clasificación de acuerdo a la cual una confiabilidad es inaceptable si es menor 0.6, es baja si está en torno de 0.7, es de moderada a elevada si va de 0.8-0.9, y es elevada si es mayor de 0.9.

Ubicación espacial para leer un párrafo.-

Fue evaluada solamente en la versión de EGRA en Nicaragua y presenta una confiabilidad inaceptable (α=0.521 para la muestra completa), explicada por el bajo número de elementos en el subtest (formado por 3 preguntas). Consideramos que este subtest presenta una confiabilidad insuficiente para ser considerado de manera independiente, lo que no impide su uso en EGRA como un criterio de control importante en los primeros grados escolares.

Nombramiento de las letras.-

Fue evaluado en ambas versiones de EGRA de manera equivalente, aunque no con los mismos ítems. En ambos casos se encuentra una confiabilidad elevada (α>0.965 para las diferentes muestras), así como evidencia de que resulta indistinto considerar ambas versiones del subtest. La alta confiabilidad puede ser explicada en parte por el gran número de elementos en el subtest (100 letras).

iIdentificación del sonido de la letra inicial de una palabra.-

S e presenta en ambas versiones de EGRA con una diferencia de apenas 2 elementos de los 10 (palabras) considerados en el subtest , y con confiabilidades similares y moderadas (α > 0.86 para las diferentes muestras).

Cuadro 3. Coeficientes Alfa de Cronbach para los diferentes subtest de EGRA en las versiones Nicaragua y Honduras

Aspecto medido			Nicaragua				Honduras			
			2º (n=2164)	3º (n=2218)	4º (n=2267)	Total (n=6649)	2º (n=615)	3º (n=597)	4º (n=526)ç	Total (n=1738)
Reading direction	RD	Ubicación espacial para leer un párrafo	0.533	0.518	0.496	0.521	NA	NA	NA	NA
Letter name recognition	LNR	Nombramiento de las letras	0.970	0.968	0.965	0.974	0.981	0.981	0.979	0.984
Recognition of initial letter sounds in words	LSR1	Identificación del sonido de la letra inicial de una palabra	0.889	0.875	0.864	0.877	0.895	0.898	0.891	0.896
Identification of words with the same initial sound	LSR2	Identificación de palabras que inician con el mismo sonido	0.676	0.768	0.786	0.770	NA	NA	NA	NA
Letter sound recall	LSR3	Recuerdo del sonido de las letras	NA	NA	NA	NA	0.922	0.929	0.926	0.927
Familiar word reading	FWR	Lectura de palabras simples	0.981	0.963	0.920	0.977	0.988	0.987	0.980	0.989
Nonsense word reading	NWR	Decodificación de palabras sin sentido	0.967	0.948	0.926	0.963	0.978	0.974	0.960	0.979
Connected text fluency	CTF	Lectura de un pasaje	SC	SC	SC	SC	SC	SC	SC	SC
Reading comprehension	RC	Comprensión de lectura de un pasaje	0.869	0.668	0.398	0.794	0.797	0.765	0.629	0.774
Oral comprehension	OC	Compresión oral de un pasaje	0.181	0.267	0.197	0.215	0.818	0.739	0.644	0.777
Writing of dictated text	WDT	Escritura de una oración	0.748	0.663	0.632	0.748	0.831	0.783	0.749	0.832

De acuerdo a Murphy y Davidshofer (1988, p. 89) una confiabilidad es inaceptable si es menor 0.6, es baja si está en torno de 0.7, es de moderada a elevada si va de 0.8-0.9 y es elevada si es mayor de 0.9.

NA: subtest no aplicado, SC: subtest aplicado pero no susceptible de hallar la confiabilidad porque no se encuentran registrados sus elementos o ítems.

Identificación de palabras que se inician con el mismo sonido.-

Fue evaluada únicamente en la versión de EGRA en Nicaragua y presentó una confiabilidad baja (α=0.77 para la muestra completa), fue aun menos confiable en el segundo grado. La baja confiabilidad puede ser explicada en parte por el bajo número de elementos del subtest (10 palabras).

Identificación o producción del sonido de las letras.-

Fue evaluado únicamente en la versión EGRA en Honduras y presenta una confiabilidad elevada (α > 0.92 para las diferentes muestras), explicada en parte por el alto número de elementos en el subtest (50 letras).

Lectura de palabras simples.-

Es un subtest igual en ambas versiones de EGRA y presentó una confiabilidad similar y elevada también en ambas (α > 0.92 para las diferentes muestras), explicada en parte por el alto número de elementos en el subtest (50 palabras).

Decodificación de palabras sin sentido.-

Está presente en ambas versiones de EGRA, con una diferencia de apenas 2 elementos de los 50 considerados en el subtest (50 palabras), y presentando confiabilidades similares y elevadas (α > 0.92 para las diferentes muestras) explicadas por el alto número de elementos del subtest (50 palabras).

Comprensión de lectura de un pasaje.-

Está presente en ambas versiones de EGRA, con diferencias en 4 de los 5 elementos del subtest (5 preguntas). Nótese que, considerando a todos los alumnos, en ambas versiones la confiabilidad es casi moderada (α= 0.79 en Nicaragua y α= 0.77 en Honduras). También en ambos decrece la confiabilidad en 3º y 4º grado en relación con 2º grado, especialmente en el 4º grado donde es inaceptable, especialmente en Nicaragua (α=0.4). Esto puede ser explicado porque el número de elementos es bajo y porque las preguntas son respondidas con alta tasa de acierto, sin presencia de variabilidad en las respuestas en los grados finales.

Comprensión oral de un pasaje.-

Está presente en ambos países, pero en versiones completamente diferentes. En Nicaragua el subtest presenta un texto de 29 palabras con 3 preguntas. En Honduras, el texto es de 51 palabras con 5 preguntas. Los resultados de confiabilidad indican que la versión de Nicaragua presenta confiabilidad inaceptable (α< 0.27 en todas las muestras), mientras que la de Honduras es casi moderada (α=0.78 en la muestra completa); en los grados superiores decae hasta ser inaceptable en el 4º grado (α=0.64).

Escritura de una oración.-

Es otro subtest de EGRA presente en ambos países, pero con versiones completamente diferentes. En la versión de Nicaragua, consta de un texto de 8 palabras con 6 criterios de evaluación. En Honduras, el texto es de 14 palabras con 7 criterios de evaluación. Los resultados de confiabilidad indican que la versión de Nicaragua presenta baja confiabilidad (α=0.75 en la muestra completa) y decae a inaceptable en 3º y 4º grado. La versión de Honduras presenta una confiabilidad moderada (α=0.83 en la muestra completa) y también se observa un decaimiento en 3º y 4º grado hasta un nivel bajo. El hecho de que las confiabilidades disminuyan cuanto mayor es el grado en ambas versiones se explica por la poca variabilidad de respuestas en el alto número de aciertos en los grados superiores.

2.1.1 Resumen

En resumen, podemos indicar que los sub-test "nombramiento de las letras", "lectura de palabras simples", "decodificación de palabras sin sentido", "recuerdo del sonido de las letras", en ambas versiones presentan elevada confiabilidad según la clasificación de Murphy y Davidshofer (1988).

Por otro lado, encontramos que tanto en "identificación del sonido de la letra inicial de una palabra" en la versión de Nicaragua como en "escritura de una oración" en la versión de Honduras presentan confiabilidad moderada.

En los casos de "identificación de palabras que inician con el mismo sonido" en Nicaragua, así como "comprensión de lectura de un pasaje" y "comprensión oral de un pasaje" en la versión de Honduras, las confiabilidades son bajas, con excepción de algunos grados: 2º grado en el primer caso y 4º grado en los dos últimos, en los que es inaceptable.

En los casos de "comprensión de lectura de un pasaje" y "escritura de una oración" en la versión de Nicaragua sólo debe ser aceptada para el 2º grado.

Finalmente, es necesario indicar que "ubicación espacial para leer un párrafo" y "comprensión oral de un pasaje" en la versión de Nicaragua, presentan una confiabilidad inaceptable para ser considerados como subtest independientes, lo cual no significa que no sean importantes para una medida general en EGRA como se reportará en la sección 2.4.

2.2 Estructura de correlaciones de los aspectos medidos de EGRA

2.2.1 Caso Nicaragua

La estructura de correlaciones entre los aspectos medidos en EGRA para el 2º, 3º y 4º grado es mostrada en el Anexo 1 y resumida en la figura 1.

Se consideran 8 aspectos medidos en EGRA, excluyendo “ubicación espacial para leer un párrafo” y “comprensión oral” que, como se reportó en la sección anterior, presentan una confiabilidad inaceptable para ser considerados como subtest. Para este análisis sí se ha considerado “lectura de un pasaje”, porque se cuenta con un puntaje en este subtest.

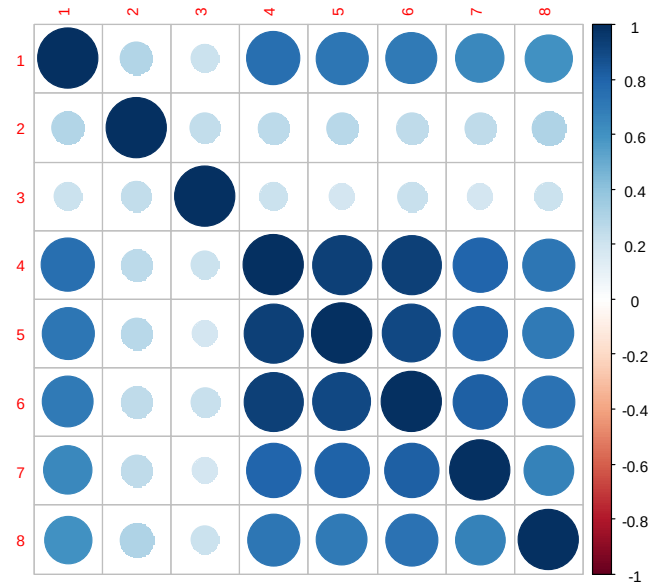
La figura 1 muestra los correspondientes ploteos de correlación (corrplot) obtenidos con el paquete Corrplot en el programa R. El tamaño de los círculos y los colores reflejan valores de correlación. Cuanto más grande y más obscuro el círculo, mayor la correlación (ver Friendly, 2002 para detalles).

En todos los grados en Nicaragua encontramos que todas las correlaciones son significativas, como se observa en el Anexo 1. Específicamente, y analizando la figura, 1 se encuentra que:

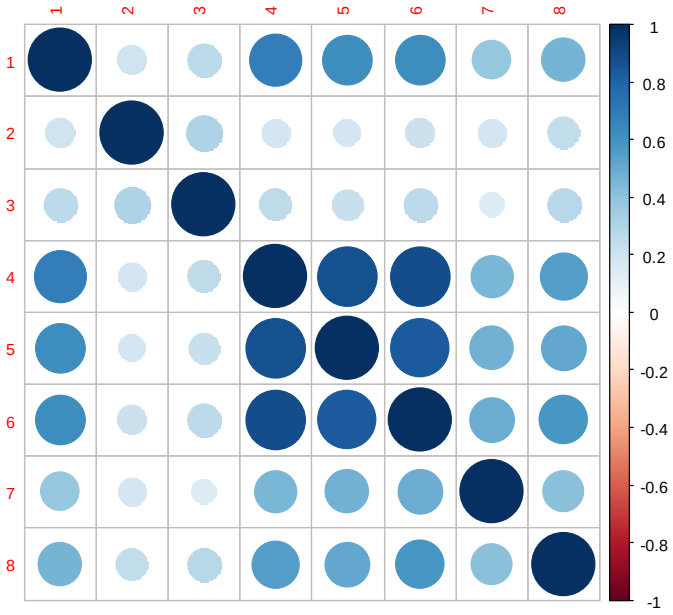
a) en los tres grados hay alta correlación entre los aspectos relacionados con tiempo, como nombramiento de las letras, lectura de palabras simples, decodificación de palabras sin sentido y lectura de un pasaje;
b) en 2º y 3er grado se encuentra una correlación mediana entre dictado y comprensión oral, y de estos con los aspectos relacionados con tiempo; en 4º grado se encuentra correlación mediana entre dictado y los aspectos relacionados con tiempo;
c) en los tres grados se encuentra correlación baja entre los aspectos fonológicos (identificación del sonido de la letra inicial de una palabra e identificación de palabras que inician con el mismo sonido) y el resto de aspectos; en el 4º grado hay una correlación baja entre comprensión de lectura de un pasaje y el resto de aspectos medidos.

Figura 1. Estructura de correlaciones entre los aspectos medidos de EGRA para el 2º, 3º y 4º grado en Nicaragua

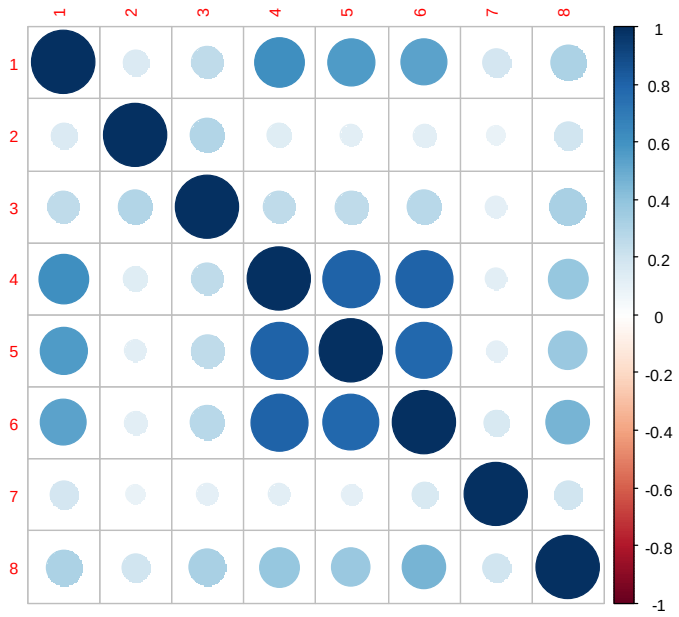
2º grado



3er grado



4º grado



1: Nombramiento de las letras, 2: Identificación del sonido de la letra inicial de una palabra, 3: Identificación de palabras que inician con el mismo sonido, 4: Lectura de palabras simples, 5: Decodificación de palabras sin sentido, 6: Lectura de un pasaje, 7: Comprensión de lectura de un pasaje, 8: Escritura de una oración

En la figura se muestran las correlaciones entre 8 aspectos medidos en EGRA. El tamaño de los círculos y los colores reflejan valores de correlación. Cuanto más grande y obscuro el círculo, mayor la correlación.

2.2.2 Caso Honduras

La estructura de correlaciones entre los aspectos medidos en EGRA para el 2º, 3º y 4º grado es mostrada en el Anexo 2 y en la figura 2. Son considerados 9 aspectos medidos en EGRA, que incluyen "lectura de un pasaje", porque se cuenta con un puntaje en este subtest.

En Honduras encontramos que en todos los grados todas las correlaciones son significativas, como se observa en el Anexo 2. Específicamente, considerando la figura 2:

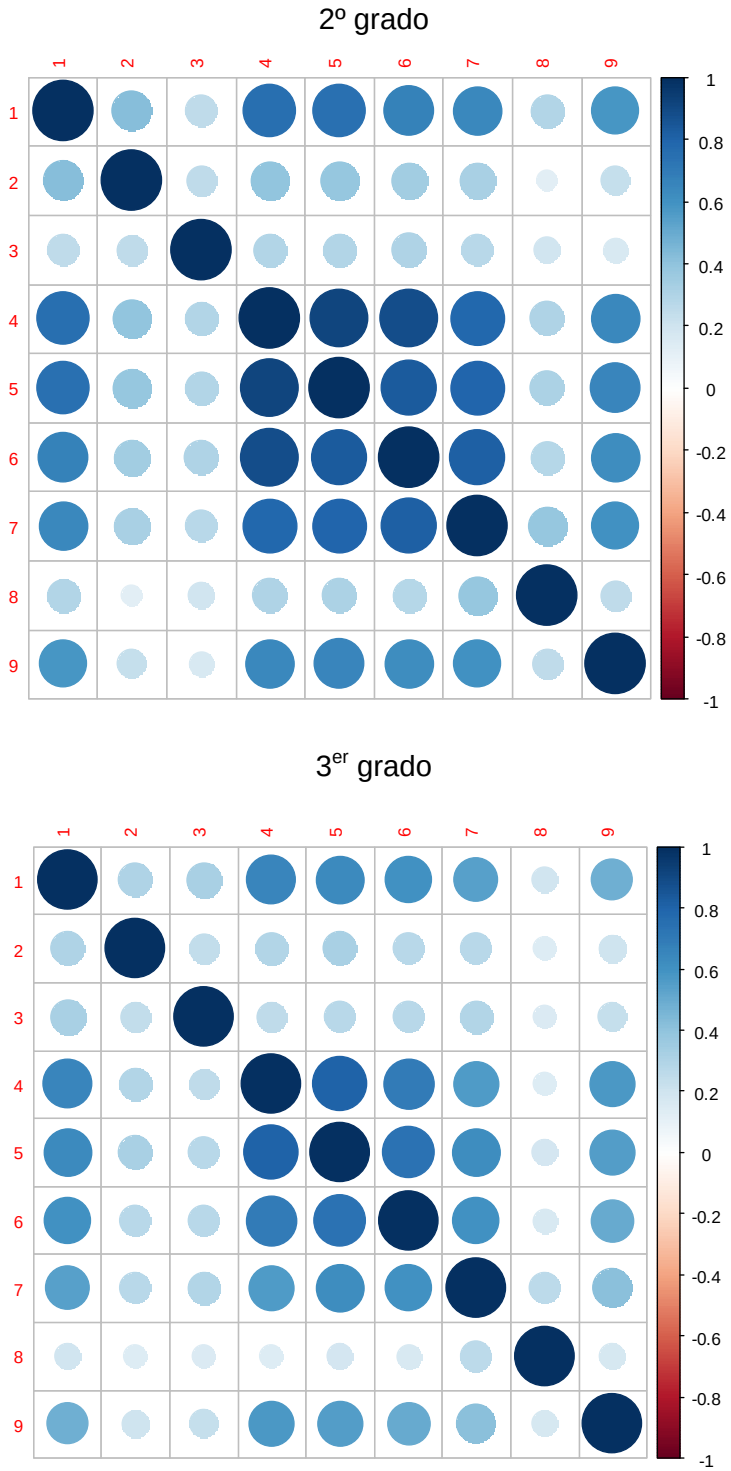
a) en los tres grados se encuentra alta correlación entre los aspectos relacionados con tiempo, (nombramiento de las letras, lectura de palabras simples, decodificación de palabras sin sentido) y lectura de un pasaje;

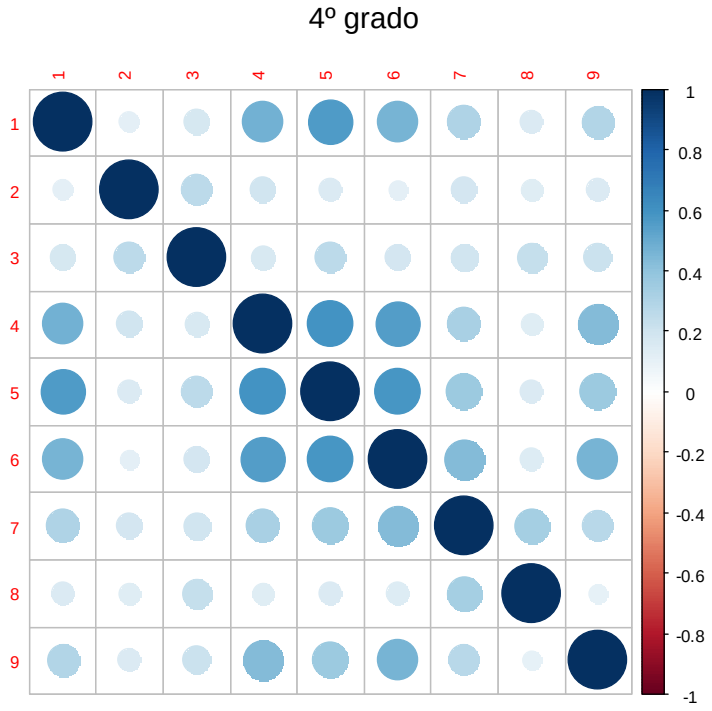
b) en 2º y 3er grado se encuentra correlación mediana entre dictado y comprensión oral, y de estos con los aspectos relacionados con tiempo, mientras que en 4º grado se encuentra correlación mediana entre dictado y los aspectos relacionados con tiempo;

c) en los tres grados se encuentra correlación baja entre los aspectos fónicos (identificación del sonido de la letra inicial de una palabra y recuerdo del sonido

de las letras) y el resto de aspectos, y en el 4º grado, una correlación baja entre comprensión de lectura de un pasaje y el resto de aspectos medidos.

Figura 2. Estructura de correlaciones entre los aspectos medidos de EGRA para el 2º, 3º y 4º grado en Honduras





1: Nombramiento de las letras, 2: Identificación del sonido de la letra inicial de una palabra, 3: Recuerdo del sonido de las letras, 4: Lectura de palabras simples, 5: Decodificación de palabras sin sentido, 6: Lectura de un pasaje, 7: Comprensión de lectura de un pasaje, 8: Compresión oral de un pasaje, 9: Escritura de una oración

En la figura 2 se muestran las correlaciones entre 8 aspectos medidos en EGRA. El tamaño de los círculos y los colores reflejan valores de correlación. Cuanto más grande y obscuro el círculo, mayor la correlación.

2.2.3 Resumen

En resumen, considerando ambas versiones, identificamos para los tres grados considerados:

Un primer grupo de subtest con alta correlación entre sí, conformado por: “nombramiento de las letras”, “lectura de palabras simples”, “decodificación de palabras sin sentido” y “lectura de un pasaje”, aspectos que son medidos en unidades por minuto.

Un segundo grupo de substest identificados para el 2º y 3er grado, conformado por “dictado” y “comprensión oral”, que presentan correlación mediana entre sí y con los aspectos del primer conjunto de subtest relacionados con tiempo; en 4º grado, la correlación es mediana entre “dictado” y los aspectos relacionados con tiempo.

Un tercer grupo conformado por los aspectos fónicos de una u otra versión, que presenta, en los tres primeros grados, baja correlación con el resto de aspectos, y en el 4º grado, una correlación baja de comprensión de lectura de un pasaje con el resto de aspectos medidos.

La estructura de correlación encontrada sugiere que EGRA puede estar midiendo más de un factor, por lo que resulta necesario hacer análisis adicionales y complementarios , los cuales son abordados en la siguiente sección.

2.3 Análisis de la estructura de EGRA

En esta sección evaluamos la estructura de EGRA tomando en cuenta todos sus subtest, tanto en la versión de Nicaragua como en la de Honduras. Los subtest son considerados aquí como elementos o "ítems", y a partir de ellos es posible obtener el Alfa de Cronbach de EGRA de manera global. Nótese que, en este caso, el Alfa de Cronbach entre subtest considera "ítems" con puntajes continuos los cuales difieren del Alfa de Cronbach dentro de cada subtest, presentado en la sección 2.1, que considera ítems con puntajes dicotómicos.

En esta sección incluimos otros índices alternativos al Alfa de Cronbach, como los índices Omega propuestos por McDonald (1995, 1999); esto debido a que como ha sido indicado en la sección 1.2.2 (Confiabilidad de pruebas), Alfa de Cronbach es una medida cuestionada en algunas situaciones (ver Revelle y Zinbarg, 2009; Sijtsma, 2009). Esto resulta especialmente importante si se considera que la estructura de correlaciones de los subtest de EGRA, reportada en la sección 2.2, nos ha dado una primera evidencia de la multidimensionalidad de EGRA, además de que Zinbarg et al. (2005) han encontrado que el Alfa de Cronbach es sensible frente a este caso, mientras que los índices Omega no lo son.

La posible evidencia de multidimensionalidad de EGRA requiere ser confirmada, por lo que en esta sección se reporta adicionalmente el análisis de los coeficientes Omega, el cual incluye, a su vez, un análisis factorial confirmatorio.

Los coeficientes Omega propuestos por McDonald (1985, 1999) son dos: Omega_t y Omega_h, y estiman, respectivamente, la saturación de factores comunes y la saturación de un factor general de un test. En este sentido, Omega_h ≤ Omega _t, obtiene la igualdad solamente cuando el test es unidimensional. Estas medidas resultan necesarias en el caso de EGRA, ya que proporcionan información complementaria al Alfa de Cronbach.

En la librería *psych* del software R (Revelle, 2008), existe una función llamada "Omega", la cual estima ambos coeficientes Omega usando un análisis factorial jerárquico, que rota los factores oblicuamente y luego realiza la transformación de Schmid-Leiman (ver Zinbarg et al., 2005). Estos análisis son reportados para ambas versiones de EGRA.

2.3.1 Caso Nicaragua

En el cuadro 4 se presenta el cálculo del Alfa de Cronbach usual y estandarizado de EGRA, así como los valores de Omega_h y Omega_t.

Cuadro 4. Alfa de Cronbach y otros índices de confiabilidad para la versión de EGRA Nicaragua (10 aspectos medidos y N=6649)

		Correlación elemento-total corregida	Alfa de Cronbach si se elimina el elemento	Alfa de Cronbach estandarizado si se elimina el elemento
RD	Reading direction	.070	.832	0.86
LNR	Letter name recognition	.733	.778	0.80
LSR1	Letter sound recognition by word	.254	.824	0.84
LSR2	Letter sound recognition between word	.380	.822	0.83
FWR	Familiar word reading	.852	.758	0.79
NWR	Nonsense word reading	.842	.779	0.79
CTF	Connected text fluency	.832	.756	0.79
RC	Reading comprehension	.660	.780	0.81
OC	Oral comprehension	.117	.851	0.86
WDT	Write dictate text	.710	.776	0.80

Alfa de Cronbach =0.816, Alfa de Cronbach estandarizado=0.84, Omega_h=0.78, Omega_t=0.88

Encontramos que el Alfa de Cronbach estandarizado de EGRA-Nicaragua es 0.84 EGRA, que puede ser considerado moderado.

Se identifican tres grupos de subtest: un primer grupo, conformado por LNR, FWR, NWR, CTF, RC y WDT, fuertemente asociados entre sí; un segundo grupo, formado por LSR1 y LSR2, moderadamente asociados, y un tercer grupo, formado por RD y OC, que no se encuentran asociados a los demás subtest y que pueden ser retirados de una medida global de EGRA.

Nótese además, que el valor de Omega_h es 0.78, que es menor que Alfa, lo que indica que la proporción de varianza de EGRA debida a un único factor general es moderada , lo cual evidencia que la homogeneidad entre los subtest de EGRA también es moderada. Por el contrario, encontramos que Omega.t es 0.88 y mayor que Alfa, lo que indica que la proporción de varianza de EGRA debido a factores comunes es casi elevada, y con ello, la consistencia interna de EGRA es adecuada.

Sobre la base de la clasificación de Zinbarg et al. (2005), la situación encontrada, donde Omega_h < Alfa < Omega_t, corresponde a una situación en la que el test es

multidimensional y las cargas factoriales en el factor general se pueden considerar iguales.

Para establecer esto con mayor claridad, en el cuadro 5 se muestran los resultados del análisis factorial realizado con la función Omega de *psych*.

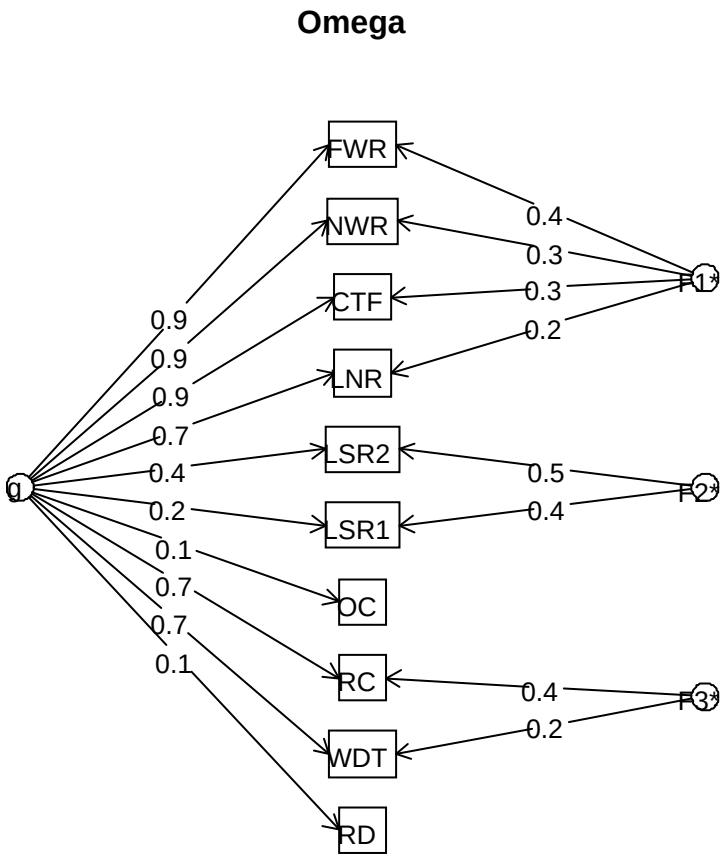
Cuadro 5. Cargas factoriales mayores que 0.2 en un análisis factorial con transformación Schmid-Leiman para los subtest de EGRA – Nicaragua

Aspectos medidos	Factor general	Factor 1*	Factor 2*	Factor 3*	Porcentaje en el factor general: H2	Porcentaje en los otros factores: u2	Porcentaje de la varianza común para cada ítem que es parte de la varianza del factor general p2
RD	0.07				0.01	0.99	0.62
LNR	0.74	0.23			0.62	0.38	0.9
LSR1	0.24		0.36		0.2	0.8	0.29
LSR2	0.37		0.54		0.43	0.57	0.32
FWR	0.9	0.36			0.94	0.06	0.86
NWR	0.87	0.33			0.87	0.13	0.88
CTF	0.89	0.31			0.89	0.11	0.89
RC	0.69			0.42	0.65	0.35	0.73
OC	0.11				0.06	0.94	0.2
WDT	0.72			0.24	0.6	0.4	0.85
valores propios	4.12	0.4	0.49	0.27			
% de varianza	41.2 %	4 %	4.9 %	2.7 %		Media	0.65
% varianza acumulada	41.2 %	45.2 %	50.1 %	52.8 %		Des. Est.	0.28
						CV	0.43

RD: Reading direction, LNR: Letter name recognition, LSR1: Letter sound recognition by word, LSR2: Letter sound recognition between word, FWR: Familiar word reading, NWR: Nonsense word reading, CTF: Connected text fluency, RC: Reading comprehension, OC; Oral comprehension, WDT: Write dictated text

De acuerdo con el cuadro 5, encontramos apenas un promedio de 0.65 en la proporción de la varianza común para cada ítem que es parte de la varianza del factor general y el porcentaje explicado es de 41.2%. Con esto se tiene una evidencia adicional de que los substest de EGRA no son homogéneos al medir un factor general. En este factor general, los subtest que presentan mayor proporción son: LNR, FWR,NWR, CTF. RC y WDT, donde los cuatro primeros forman un primer factor (F1) y los dos últimos otro factor (F3). LSR1 y LSR2 presentan una proporción baja en el factor general y constituyen también un factor propio (F2). Finalmente, RD y OC presentan una proporción menor a 0.2 en el factor general, así como en los otros factores, sin ser parte de ninguno de ellos, por lo que se recomienda su exclusión para una medida global de EGRA-Nicaragua. Esta estructura puede ser apreciada en la figura 3.

Figura 3. Estructura de EGRA-Nicaragua considerando todos los subtest basados en un análisis factorial (solución de *minimum residual OLS*) con 3 factores con rotación *"oblimin"* usando la transformación Schmid-Leiman obtenida en *Psych Library*



RD: Reading direction, LNR: Letter name recognition, LSR1: Letter sound recognition by word, LSR2: Letter sound recognition between word, FWR: Familiar word reading, NWR: Nonsense word reading, CTF: Connected text fluency, RC: Reading comprehension, OC; Oral comprehension, WDT: Write dictated text

Como se ha establecido antes, se confirma, con la figura mostrada, que RD y OC son aspectos medidos inadecuadamente en la versión de Nicaragua, ya que no proporcionan ninguna evidencia a favor de una medida común de EGRA.

Adicionalmente, se distinguen tres factores comunes: un primer factor conformado por FWR, NWR, CTF y LNR, aspectos relacionados con precisión o fluidez de palabras familiares, pseudopalabras, palabras conectadas y letras en una unidad de tiempo, aspectos relacionados con automatización de lectura. Un segundo factor está conformado por LSR1 y LSR2, relacionados con porcentajes de logro en aspectos específicos de conciencia fonética y fonológica. Un tercer factor está conformado por RC y WDT, aspectos también relacionados con porcentajes de logro en tareas de comprensión y dictado.

Para confirmar si la estructura de tres factores o la estructura de un solo factor general se ajusta a la muestra de Nicaragua, se realizó un análisis factorial confirmatorio dentro del contexto del modelamiento de ecuaciones estructurales (Kline, 2005; Reisinger y Mavondo, 2006), cuyos índices de ajuste son presentados en el cuadro 6:

Cuadro 6. Índices de ajuste del análisis factorial confirmatorio de EGRA-Nicaragua para un modelo de tres factores y un modelo de un factor general (N=6649)

	Adecuación para el modelo de tres factores	Adecuación de ajuste para un factor general y no otro grupo de factores
F: Minimized fitting criterion	0.04	0.69
Degrees of freedom	18	35
Test Chi square	269.09 with prob< 1.1e-46	4567.3 with prob < 0
Root Mean Square Error of Approximation (RMSEA)	0.046	0.1
Bayesian Information Criterion (BIC)	110.64	4259.22

Los resultados permiten observar que, tanto el modelo de tres factores propuestos, como el de un único factor general, presentan un valor en el estadístico chi-cuadrado, que es significativo; es decir, los datos observados se diferencian de forma marcada de los datos que se esperaría obtener si el modelo fuese adecuado.

Podríamos pensar que el modelo de 3 factores, así como el modelo de un factor general, no funcionan adecuadamente en la muestra. Sin embargo, se sabe que cuando el tamaño de muestra es grande, la estadística chi-cuadrado siempre es significativa; de allí que sea importante considerar otros criterios de ajuste.

Un criterio de ajuste más adecuado que la estadística Chi-cuadrado es el índice RMSEA desarrollado por (Steiger, 1990). Un ajuste puede ser considerado bueno si el valor de RMSEA es menor de 0.6 (según Hu y Bentler, 1999) o incluso menor de 0.07 (según Steiger, 2007).

En el caso del modelo de 3 factores, encontramos un valor de RMSEA de 0.046, que indica que el modelo es bueno, en contraste con el modelo con un solo factor general que presenta RMSEA de 0.10.

Otra medida de ajuste reportada en el cuadro es el valor del Bayesian Information Criterion (BIC). BIC no es una medida absoluta, por lo que se usa para comparar el ajuste de modelos alternativos para los mismos datos, donde un modelo con menor BIC indica que es el mejor modelo de los que están en comparación. Para el modelo de tres factores el valor de BIC es 110.64, mientras que para el modelo de un único factor general el valor de BIC es 4259.22, lo que indica claramente que el modelo de tres factores es el mejor modelo para explicar la estructura de EGRA en la versión de Nicaragua.

2.3.2 Honduras

En el cuadro 7 se presenta el cálculo deAlfa de Cronbach usual y estandarizado de EGRA así como los valores de Omega_h y Omega_t.

Cuadro 7. Alfa de Cronbach y otros índices de confiabilidad para la versión de EGRA-Honduras (10 aspectos medidos y N=1738)

		Correlación elemento-total corregida	Alfa de Cronbach si se elimina el elemento	Alfa de Cronbach estandarizado si se elimina el elemento
LNR	Letter name recognition	.696	.805	0.85
LSR1	Letter sound recognition by word	.280	.843	0.88
LSR3	Letter sound recall	.326	.847	0.88
FWR	Familiar word reading	.773	.793	0.85
NWR	Nonsense word reading	.804	.808	0.84
CTF	Connected text fluency	.754	.804	0.85
RC	Reading comprehension	.664	.806	0.85
OC	Oral comprehension	.327	.847	0.88
WDT	Write dictate text	.692	.807	0.86

Alfa de Cronbach=0.837, Alfa de Cronbach estandarizado=0.88, Omega_h=0.79, Omega_t=0.90

Encontramos que el Alfa de Cronbach estandarizado de EGRA-Honduras es 0.88 EGRA, que puede ser considerado casi elevado.

Se identifican dos grupos de subtest: un primer grupo, conformado por LNR, FWR, NWR, CTF, RC y WDT, que se encuentran fuertemente asociados entre sí; un segundo grupo, formado por LSR1, LSR3 y OC, moderadamente asociados a los otros subtest de EGRA.

Nótese además que el valor de Omega_h es 0.79, menor que Alfa, lo que indica que la proporción de varianza de EGRA debida a un único factor general es moderada, evidencia de que la homogeneidad entre los subtest de EGRA también es moderada. Por el contrario, encontramos que Omega.t es 0.90 y mayor que Alfa, lo que indica que la proporción de varianza de EGRA debido a factores comunes es elevada y, con ello, la consistencia interna de EGRA es adecuada.

Sobre la base de la clasificación de Zinbarg et al. (2005), la situación encontrada, donde Omega_h < Alfa < Omega_t, corresponde a una situación en la que el test es multidimensional y las cargas factoriales en el factor general se pueden considerar iguales.

Para establecer esto con mayor claridad, en el cuadro 8 se muestran los resultados del análisis factorial realizado con la función Omega de *psych*.

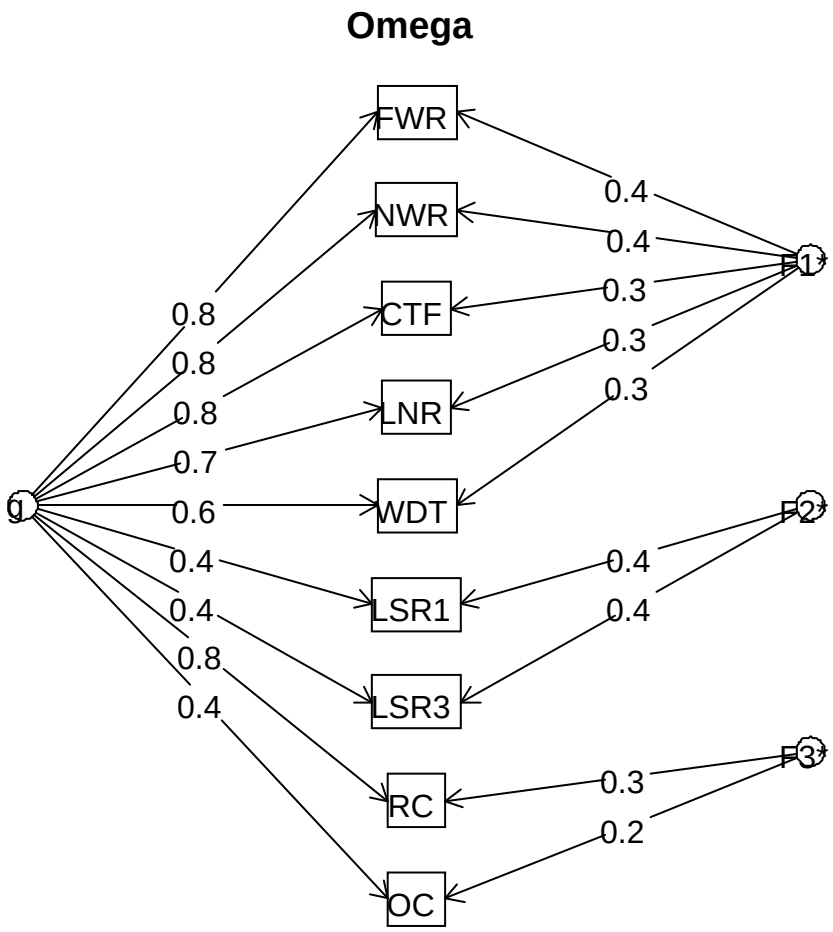
Cuadro 8. Cargas factoriales mayors que 0.2 en un análisis factorial con transformación Schmid-Leiman para los subtest de EGRA-Honduras

	Factor general	Factor 1*	Factor 2*	Factor 3*	Porcentaje in el factor general: H2	Porcentaje en los otros factores: u2	Porcentaje de la varianza comun para cada ítem que es parte de la varianza del factor general: p2
LNR	0.73	0.3			0.63	0.37	0.84
LSR1	0.36		0.38		0.28	0.72	0.46
LSR3	0.35		0.36		0.26	0.74	0.48
FWR	0.79	0.43			0.81	0.19	0.77
NWR	0.83	0.37			0.82	0.18	0.83
CTF	0.8	0.32			0.75	0.25	0.85
RC	0.79			0.35	0.75	0.25	0.83

	Factor general	Factor 1*	Factor 2*	Factor 3*	Porcentaje in el factor general: H2	Porcentaje en los otros factores: u2	Porcentaje de la varianza comun para cada ítem que es parte de la varianza del factor general: p2
OC	0.4			0.23	0.24	0.76	0.67
WDT	0.63	0.28			0.47	0.53	0.83
Valores propios	3.9	0.6	0.32	0.19			
% de varianza	39.0 %	6 %	3.2 %	1.9 %		mean	0.73
% de varianza acumulada	39.0 %	45.0 %	48.2 %	51.2 %		sd	0.16
						CV	0.22

De acuerdo con el cuadro 8, encontramos apenas un promedio de 0.73 en la proporción de la varianza común para cada ítem que es varianza del factor general y que el porcentaje explicado es de 39%. Esto constituye evidencia adicional de que los substest de EGRA no son homogéneos al medir un factor general. En este factor general, los subtest que presentan mayor proporción son: LNR, FWR,NWR, CTF, WDT y RC, donde los cinco primeros forman un primer factor (F1), y el último forma parte de otro factor junto a OC (F3). Por otro lado, LSR1 y LSR2 presentan una proporción baja en el factor general y constituyen también un factor propio (F2). En este caso, a diferencia de EGRA-Nicaragua, todos los subtest conforman los diferentes factores identificados. Esta estructura puede ser apreciada en la figura 4.

Figura 4. Estructura de EGRA-Honduras considerando todos los subtest, basada en un análisis factorial (*minimum residual OLS solution*) con 3 factores con rotación *"oblimin"* usando la transformación Schmid-Leiman obtenida en *Psych Library*



LNR: Letter name recognition, LSR1: Letter sound recognition by word, LSR2: Letter sound recognition between word, FWR: Familiar word reading, NWR: Nonsense word reading, CTF: Connected text fluency, RC: Reading comprehension, OC; Oral comprehension, WDT: Write dictate text

Como se ha establecido antes, se confirma con la figura 4 que todos los aspectos medidos en EGRA para la versión de Nicaragua son adecuados, pero no proporcionan ninguna evidencia a favor de una medida común o factor general de EGRA.

Adicionalmente, se distinguen tres factores comunes en EGRA: un primer factor, conformado por FWR, NWR, CTF, LNR y WDT, aspectos relacionados con precisión o fluidez de palabras familiares, pseudopalabras, palabras conectadas y letras en una unidad de tiempo, junto al porcentaje de logro en un dictado, relacionados con automatización de lectura. Un segundo factor está conformado por LSR1 y LSR3, relacionados con porcentajes de logro en aspectos específicos de conciencia fonética y fonológica. Un

tercer factor está conformado por RC y OC, aspectos también relacionados con porcentajes de logro en tareas de comprensión.

Para confirmar si la estructura de tres factores o la estructura de un solo factor general se ajusta a la muestra de Honduras, se realizó un análisis factorial confirmatorio dentro del contexto del modelamiento de ecuaciones estructurales (Kline, 2005, Reisinger y Mavondo, 2006), cuyos índices de ajuste son presentados en el cuadro 9:

Cuadro 9. Índices de ajuste del análisis factorial confirmatorio de EGRA-Nicaragua para un modelo de tres factores y un modelo de un factor general (N=1738)

	Adecuación para un modelo de tres factores	Adecuación de justo un factor general y no grupo de factores
F: Minimized fitting criterion	0.02	0.4
Degrees of freedom	12	27
Test Chi square	39.55 with prob< 8.6e-5	697.97 with prob < 1e-129
Root Mean Square Error of Approximation (RMSEA)	0.036	0.119
Bayesian Information Criterion (BIC)	-49.98	496.53

Como ocurrió con la versión EGRA-Honduras, los resultados que consideran la estadística Chi-cuadrado presentan un resultado significativo, es decir, que el modelo de 3 factores, así como el modelo de un factor general, no funcionan adecuadamente en la muestra. Sin embargo, como se ha comentado antes, esta estadística no es confiable debido al gran tamaño de la muestra, por lo que debemos considerar otros criterios.

En el caso del modelo de 3 factores, encontramos un valor de RMSEA de 0.036, que indica que el modelo es bueno, en contraste con el modelo con un solo factor general que presenta RMSEA de 0.11. El valor de BIC para este modelo es también 49.98, mientras que para el modelo de un único factor general es 496.53, lo que indica claramente que el modelo de tres factores es el mejor para explicar la estructura de EGRA en la versión de Honduras.

2.3.3 Resumen

En resumen, considerando ambas versiones, para la muestra completa en los tres grados considerados, identificamos:

Un primer factor conformado por los subtest “nombramiento de las letras”, “lectura de palabras simples”, “decodificación de palabras sin sentido” y “lectura de un pasaje”, aspectos que son medidos en unidades por minuto.

Un segundo factor conformado por los substest relacionados con fonética y fonología: la identificación de sonidos de letras iniciales de una palabra y de palabras que inician con el mismo sonido junto al recuerdo de sonidos de letras; los dos primeros en la versión de Nicaragua, y el primero y tercero en la prueba de Honduras.

Un tercer factor está conformado por comprensión de lectura, que se encuentra asociado con el dictado de un texto en el caso de Nicaragua, y con comprensión oral en el caso de Honduras. Este factor no resulta claro, porque los dos subtest mencionados no presentan una confiabilidad alta y requieren ser mejorados en futuras aplicaciones de EGRA.

También en la versión EGRA-Nicaragua queda claro que RD y OC son dos subtest que no deben ser considerados para análisis complementarios.

La estructura de factores encontrada sugiere que EGRA es multidimensional (automatización y precisión para la lectura, conciencia fónica y fonológica y comprensión). EGRA presenta alta consistencia interna y homogeneidad moderada.

2.4 Análisis de la validez concurrente de EGRA

En los cuadros 10 y 11 se presenta la correlación existente entre los diferentes aspectos medidos en EGRA y las pruebas de español de Nicaragua y Honduras, respectivamente. En el caso de las pruebas de español, se consideran los puntajes fila (número de aciertos en la prueba) y escala (score en un modelo de teoría de respuesta al ítem considerado para cada prueba).

Este análisis está restringido a las muestras en que los estudiantes dieron ambas pruebas. En el caso de Nicaragua esto únicamente ocurrió en el 4º grado, y en el caso de Honduras, en 2º, 3º y 4º grado.

Además, el análisis está restringido a los aspectos o subtest con adecuada medición en EGRA-Nicaragua y EGRA-Honduras.

2.4.1 Nicaragua

Cuadro 10. Correlación de Pearson entre aspectos medidos en EGRA y puntaje fila y en escala para la prueba de español en una muestra de n=374 estudiantes de Nicaragua del 4º grado

Aspectos medidos EGRA	Puntaje fila en español			Puntaje en escala en español		
	R	Sig		R	Sig	
Puntaje en escala en español	0.987	0.000	**	1		
Nombramiento de las letras	0.352	0.000	**	0.366	0.000	**
Identificación del sonido de la letra inicial de una palabra	0.268	0.000	**	0.276	0.000	**
Identificación de palabras que inician con el mismo sonido	0.303	0.000	**	0.306	0.000	**
Lectura de palabras simples	0.310	0.000	**	0.313	0.000	**
Decodificación de palabras sin sentido	0.313	0.000	**	0.315	0.000	**

Aspectos medidos EGRA	Puntaje fila en español			Puntaje en escala en español		
	R	Sig		R	Sig	
Lectura de un pasaje	0.407	0.000	**	0.413	0.000	**
Comprensión de lectura de un pasaje	0.208	0.000	**	0.214	0.000	
Escritura de una oración	0.360	0.000	**	0.361	0.000	**

En Nicaragua, los aspectos medidos de EGRA se encuentran correlacionados de manera significativa con los puntajes de español, con coeficientes ligeramente mejores en el caso de puntajes de escala. Los coeficientes, en orden de mayor a menor, son "lectura de un pasaje", "escritura de una oración", "nombramiento de las letras", "decodificación de palabras sin sentido", "lectura de palabras simples", "identificación de palabras que inician con el mismo sonido", "identificación del sonido de la letra inicial de una palabra" y finalmente, "comprensión de lectura de un pasaje".

2.4.2 Honduras

Cuadro 11. Correlación de Pearson entre aspectos medidos en EGRA, puntaje fila y de escala de la prueba de español en estudiantes de Honduras (n=262 en 2º, n=213 en 3er, n=265 en 4º grado)

2º grado

Aspectos medidos	Puntaje fila en español			Puntaje en escala en español		
	R	Sig		R	Sig	
Puntaje en escala en español	0.988	0.000		1		
Nombramiento de las letras	0.220	0.000	**	0.204	0.001	**
Recuerdo del sonido de las letras	0.041	0.515		0.040	0.520	
Identificación del sonido de la letra inicial de una palabra	0.094	0.128		0.075	0.224	
Lectura de palabras simples	0.181	0.003	**	0.162	0.009	**
Decodificación de palabras sin sentido	0.192	0.002	**	0.177	0.004	**
Lectura de un pasaje	0.189	0.002	**	0.168	0.006	**
Comprensión de lectura de un pasaje	0.229	0.000	**	0.220	0.000	**
Compresión oral de un pasaje	0.136	0.027	*	0.130	0.036	*
Escritura de una oración	0.303	0.000	**	0.267	0.000	**

3er grado

Aspectos medidos	Puntaje fila en español			Puntaje en escala en español		
	R	Sig		R	Sig	
Puntaje en escala en español	0.980	0.000		1		
Nombramiento de las letras	0.390	0.000	**	0.375	0.000	**
Recuerdo del sonido de las letras	0.282	0.000	**	0.266	0.000	*
Identificación del sonido de la letra inicial de una palabra	0.262	0.000	**	0.262	0.000	**
Lectura de palabras simples	0.418	0.000	**	0.395	0.000	**
Decodificación de palabras sin sentido	0.442	0.000	**	0.415	0.000	**

Aspectos medidos	Puntaje fila en español			Puntaje en escala en español		
	R	Sig		R	Sig	
Lectura de un pasaje	0.443	0.000	**	0.418	0.000	**
Comprensión de lectura de un pasaje	0.377	0.000	**	0.336	0.000	**
Compresión oral de un pasaje	0.175	0.011	*	0.161	0.019	*
Escritura de una oración	0.341	0.000	**	0.321	0.000	**

4º grado

Aspectos medidos	Puntaje fila en español			Puntaje en escala en español		
	R	Sig		R	Sig	
Puntaje en escala en español	0.987	0.000		1		
Nombramiento de las letras	0.283	0.000	**	0.274	0.000	**
Recuerdo del sonido de las letras	0.155	0.012	*	0.148	0.017	*
Identificación del sonido de la letra inicial de una palabra	0.289	0.000	**	0.283	0.000	**
Lectura de palabras simples	0.279	0.000	**	0.264	0.000	**
Decodificación de palabras sin sentido	0.248	0.000	**	0.232	0.000	**
Lectura de un pasaje	0.373	0.000	**	0.354	0.000	**
Comprensión de lectura de un pasaje	0.263	0.000	**	0.240	0.000	**
Compresión oral de un pasaje	0.257	0.000	*	0.262	0.000	*
Escritura de una oración	0.378	0.000	**	0.381	0.000	**

En el caso de Honduras, en el 2º grado se encontraron correlaciones significativas de las pruebas de español con los aspectos medidos en EGRA, con excepción de recuerdo del sonido de las letras e identificación del sonido de la letra inicial de una palabra. Entre las correlaciones significativas, el mayor coeficiente es obtenido con escritura de una oración y el menor, con comprensión oral de un pasaje.

En el 3er y 4º grado se encontraron correlaciones significativas de las pruebas de español con todos los aspectos medidos de EGRA. Entre las correlaciones significativas, en el 3er grado el mayor coeficiente es obtenido por "lectura de un pasaje" y el menor, con "comprensión oral de un pasaje". En el 4º grado, el mayor coeficiente es obtenido por "escritura de una oración" y el menor, por "recuerdo del sonido de las letras".

2.4.3 Resumen

En Nicaragua los diferentes aspectos medidos de EGRA se encuentran significativamente relacionados con la prueba de español dada por los estudiantes de 4º grado.

En Honduras, los aspectos medidos de EGRA: "nombramiento de las letras", "lectura de palabras simples", "decodificación de palabras sin sentido", "lectura de un pasaje", "comprensión de lectura de un pasaje", "escritura de una oración", se encuentran significativamente relacionados con la prueba de español en los diferentes grados. Y "recuerdo del sonido de las letras" e "identificación del sonido de la letra inicial de una palabra" se encuentran significativamente relacionados solamente en el tercer y cuarto

grado. Finalmente, “comprensión oral” se encuentra significativamente relacionado, en los diferentes grados, con la prueba de español.

Tomando en cuenta estos resultados podemos considerar que EGRA se encuentra, en alguna medida, asociada con el desempeño en español en los grados de 2º, 3ºo y 4º tanto en Nicaragua como en Honduras.

Capítulo 3. Conclusiones y recomendaciones

Como resumen de nuestro análisis de los conjuntos de datos de las evaluaciones de Nicaragua y Honduras, podemos ofrecer lo siguiente:

Análisis autónomo de EGRA:

- Utilizando medidas estándar, encontramos un alto nivel de fiabilidad entre las mismas sub-pruebas de EGRA. Las sub-pruebas que obtuvieron los peores resultados en las medidas de fiabilidad (como aquella orientada a la comprensión)- en parte por una similaridad generalizada en los resultados de los estudiantes—también fueron considerados como los elementos menos críticos del EGRA en cuanto a la validez de medición.
- La fiabilidad entre las sub-pruebas de EGRA también fue alta en ambos países, en la mayoría de los grados y para la mayoría de las sub-pruebas. Como era de esperarse, las correlaciones entre los ítems relacionados con la fluidez y entre los ítems de fluidez y comprensión fue bastante alta (pero solo antes del grado 4).
- Tanto en Nicaragua como en Honduras, las habilidades en general de la lectura oral fueron razonablemente buenas una vez llegado al grado 4, cosa que tuvo un efecto adverso debilitando la estructura de correlación entre las sub-pruebas de EGRA a ese nivel y confirmando que EGRA es más apropiado para los grados *iniciales* en estos países. La tendencia del EGRA a tener una performance menos confiable en los grados más avanzados podría ser mitigada de usar una versión más difícil del EGRA de la que se utilizó en estos casos.
- Se encontró que la consistencia interna (en la forma de valores alfa y omega) fue moderadamente alta si se excluían las sub-pruebas afectadas por la falta de variabilidad en los resultados y la adquisición de habilidades hasta el grado 4.

Validez concurrente de EGRA y las pruebas escritas:

- Dado que el EGRA y las pruebas escritas no fueron diseñados para ser comparados, compensamos evaluando la correlación entre las sub-pruebas claves del EGRA— concretamente la fluidez en la lectura narrativa y comprensión de lectura—y el puntaje general en la prueba escrita.
- Para la fluidez en la lectura narrativa, las correlaciones con los puntajes generales de las pruebas de comprensión tuvieron un promedio de 0.35 (o 0.4 si un aparente caso anómalo del grado 2 de Honduras fuese excluido).
- La correlación entre las sub-pruebas de *fluidez* de EGRA y la evaluación general de comprensión fue tan buena como la correlación promedio entre cualquiera de ítems de comprensión de EGRA y la evaluación general de comprensión—o sea, alrededor de 0.5.

- Sería útil una mayor investigación para comparar algunas sub-tareas claves de EGRA con sus correspondientes ítems individuales en una prueba escrita de comprensión.

La conclusión de este estudio en particular es que una evaluación enfocada en las capacidades orales relacionadas a la fluidez es de moderada a altamente fiable (dependiendo de la sub-prueba en cuestión, siendo los de fluidez generalmente más fiables) y moderadamente bien correlacionada con las medidas de comprensión. Este hallazgo, conjuntamente con la evidencia que proviene de países desarrollados, es alentador con respecto a la aplicación de evaluaciones orales como precursor o proxy para habilidades más avanzadas. Se deben de llevar a cabo estudios adicionales sobre las propiedades de estas evaluaciones, tanto para mejorarlos como para aumentar la facilidad de su uso, utilizando métodos que estén diseñados específicamente para este propósito.

Bibliografía

Abadzi, H. (2010). *Reading fluency measurements in EFA FTI partner countries: Outcomes and improvement prospects.* Working Paper Series, Education For All, Fast Track Initiative. Disponible en < http://www.educationfasttrack.org/media/ReadingFluencyupdated.pdf> (accesado el 20 de noviembre 2010).

Abedi, J. (1997). *Dimensionality of NAEP subscale scores in mathematics.* CSE Technical Report 428. Disponible en <http://www.cse.ucla.edu/Reports/TECH428.pdf> (accesado el 10 de mayo 2010).

Agencia de los Estados Unidos para el Desarrollo Internacional (USAID), EQUIP1. (2008). *Estudio anual 2007: rendimiento académico de los estudiantes de 3er grado en español y matemáticas en las escuelas validando el nuevo currículo.* Washington, DC: American Institutes for Research (AIR).

American Educational Research Association (AERA), American Psychological Association (APA) y National Council on Measurement in Education (NCME). (1999). *Standards for educational and psychological testing.* Preparado por un comité conjunto de la AERA, APA, y NCME. Washington, DC: AERA.

Attali, Y. (2005). Reliability of speeded number-right multiple-choice tests. *Applied Psychological Measurement, 29,* 357-368.

Bailey, E J., Bricker, D. (1986). A psychometric study of a criterion-referenced assessment instrument designed for infants and young children. *Journal of Early Intervention, 10*(2): 124-134.

Baker, F. B., Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker.

Bazán, J., Millones, J. (2002a). Evaluación psicométrica de las pruebas CRECER 98. En J. Rodriguez y S. Vargas (Eds), *Análisis de los resultados y metodología de las pruebas CRECER 1998* (pp. 171-195). Documento de trabajo 13. Lima: MECEP-Ministerio de Educación. Disponible en <http://www.minedu.gob.pe/umc/publicaciones/mecep/doc13/13i.pdf > (accesado el 20 de mayo 2010).

Bazán, J., Millones, J. (2002b). Evaluación psicométrica de las preguntas de las pruebas CRECER 98. En J. Rodriguez y S. Vargas (Eds), *Análisis de los resultados y metodología de las pruebas CRECER 1998* (pp. 141-170). Documento de trabajo 13. Lima: MECEP-Ministerio de Educación. Disponible en <http://www.minedu.gob.pe/umc/publicaciones/mecep/doc13/13h.pdf >(accesado el 20 de mayo 2010).

Bond, T. G., Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Bonnet, G. (2006). Tener presentes las singularidades lingüísticas y culturales en las evaluaciones internacionales de las competencias de los alumnos: ¿una nueva dimensión para PISA? *Revista de Educación*, extraordinario 2006, pp. 91-109. Disponible en <http://www.ince.mec.es/revedu/extra2006.pdf> (accesado el 20 mayo 2010).

Bradlow, E. T., Wainer, H., Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64,* 153-168.

Burga, A. (2005, noviembre). *La unidimensionalidad de un instrument de medición: Perspectiva factorial.* Lima, Perú: Unidad de Medición de la Calidad (UMC), Ministerio de Educación. Disponible en http://www2.minedu.gob.pe/umc/admin/images/publicaciones/artiumc/2.pdf

Carmines, E. G., Zeller, R. A. (1979). *Reliability and validity assessment.* Beverly Hills, California: Sage.

Christensen, K. B. (2005). A Monte Carlo approach to unidimensionality testing in polytomous Rasch models. December 15, 2005. Disponible en <http://pubhealth.ku.dk/bs/publikationer/rr-06-4.pdf/> (accesado el 20 de mayo 2010).

Christensen, K. B., Bjorner, J. B. (2003). *SAS macros for Rasch based latent variable modelling* (Tech. Rep. No. 03/13). Department of Biostatistics, University of Copenhagen. Disponible de <http://pubhealth.ku.dk/bs/publikationer>) (accesado el 20 de mayo 2010).

Christensen, K. B., Bjorner, J. B., Kreiner, S., Petersen, J. H. (2002). Testing unidimensionality in polytomous Rasch models. *Psychometrika*, *67* (4), 563-574.

Centro de Investigación y Acción Educativa Social (CIASES). (2010). *Honduras diagnóstico de capacidades de lectura: informe sobre resultados del estudio en escuelas PROHECO.* Preparado para el Banco Mundial. Research Triangle Park, North Carolina: RTI. Disponible en < https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&ID=263 > (accesado el 20 de julio 2010)

Centro de Investigación y Acción Educativa Social (CIASES) y RTI International (2009). *Informe de resultados: EGRA 2008.* Preparado para USAID, proyecto Education Data for Decision Making (EdData II), Orden de Trabajo nº 5. Research Triangle Park, North Carolina: RTI. Disponible en < https://www.eddataglobal.org/documents/index.cfm?fuseaction=pubDetail&ID=198> (accesado el 20 de agosto 2010)

Crocker, L., Algina, J. (1986). Factors that affect reliability coefficients. En L. Crocker y J. Algina (Eds.), *Introduction to classical and modern test theory* (pp. 143-146). Fort Worth, Texas: Harcourt.

Friendly, M. (2002). Corrgrams: Exploratory displays for correlation matrices. *The American Statistician, 56,* 316–324.

Gove, A., Cvelich, P. (2010). *Early reading: Igniting education for all.* Informe por el Early Grade Reading Community of Practice. Research Triangle Park, North Carolina: RTI Inernational. Disponible en <http://www.rti.org/pubs/early-reading-report_gove_cvelich.pdf> (accesado el 10 de diciembre 2010)

Haertel, E. (1985). Construct validity and criterion-referenced testing. *Review of Educational Research, 55*(1), 23-46.

Hu, L. T., Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6* (1), 1-55.

Joint Committee on Standards for Educational Evaluation. (2003). *The student evaluation standards: How to improve evaluations of students.* Newbury Park, California: Corwin Press. Disponible en < http://www.wmich.edu/evalctr/jc/briefing/ses/> (accesado el 20 de agosto 2010)

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford Press.

Knol, D. L., Berger, M. P. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26,* 457-477.

Kolen, M. J., Brennan R. L. (2004). *Test equating, scaling, and linking.* New York: Springer.

Kudo, I., Bazan, J. (2009). *Measuring beginner reading skills: An empirical evaluation of alternative instruments and their potential use for policymaking and accountability in Peru.* Documento de trabajo de investigación sobre políticas, 4812. Washington, DC: Banco Mundial.

Lord, F. M., Novick, M. R. (1968). *Statistical theories of mental test s*cores. Reading, Massachusetts: Addison-Wesley.

Martin, M., Kelly, D. (Eds.). (1998). *TIMSS technical report volume III: Implementation and analysis, final year of secondary school (Population 3).* Third International Mathematics and Science Study (TIMSS). Chestnut Hill, Massachusetts: Boston College. Disponible en < http://timss.bc.edu/timss1995i/TIMSSPDF/TR3book.pdf> (accesado el 20 de agosto 2010)

Martinez, R. (2006). La metodología de los estudios PISA. *Revista de Educación*, extraordinario 2006, pp. 111-129. Disponible en <http://www.ince.mec.es/revedu/extra2006.pdf> (accesado el 20 de junio 2010)

McDonald, R. P. (1985). *Factor analysis and related methods.* Hillsdale, New Jersey: Erlbaum.

McDonald, R. P. (1999). *Test theory: A unified treatment.* Mahwah, New Jersey: Erlbaum.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika, 56,* 177-196.

Muñiz, J. (1990). *Teoría de respuesta a los ítems: un nuevo enfoque en la evolución psicológica y educativa.* Madrid: Ediciones Pirámide, S.A.

Murphy, K. R., Davidshofer, C. O. (1988). *Psychological testing: Principles and applications.* Englewood Cliffs, New Jersey: Prentice Hall.

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

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.

Programme for International Student Assessment (PISA). (2005). *PISA 2003 technical report.* Disponible en <http://www.oecd.org/dataoecd/49/60/35188570.pdf> (accesado el 20 de mayo 2010).

Reckase, M. D. (1985) The difficulty of items that measure more than one ability. *Applied Psychological Measurement, 9,* 401-412.

Reisinger, Y., Mavondo, F. (2006). Structural equation modeling: Critical issues and new developments. *Journal of Travel and Tourism Marketing, 21*(4), 41-71.

Revelle, W. (2008). *Psych. Procedures for personality and psychological research.* R package version 1.0-51.

Revelle, W., Zinbarg, R. (2009). Coefficients alpha, beta, Omega, and the glb [greatest lowest bound]: Comments on Sijtsma. *Psychometrika, 74,* 145-154.

Roskos, K., D. Strickland, J. Haase, S. Malik. (2009). *First principles for early grades reading programs in developing countries.* Preparado para USAID, Education Quality Improvement Program (EQUIP1). Washington, DC: American Institutes for Research (AIR). Disponible en < http://www.equip123.net/docs/e1-EarlyGradesToolkit.pdf> (accesado el 20 de agosto 2010)

RTI International. (2009). *Manual para la evaluación inicial de la lectura en niños de educación primaria.* Preparado para USAID, proyecto do Education Data for Decision Making (EdData II), Orden de Trabajo nº 3. Research Triangle Park, North Carolina: RTI. Disponible en < http://pdf.usaid.gov/pdf_docs/PNADS441.pdf> (accesado el 20 de julio 2010)

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74,*107-120.

Steiger, J. H. (1990). Structural model evaluation and modification. *Multivariate Behavioral Research, 25,* 214-212.

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences, 42*(5), 893-898.

Swaminathan, H., Hambleton, R., Algina, J. (2005). Reliability of criterion-referenced test: A decision-theoretic formulation. *Journal of Educational Measurement, 11,* 263-267.

Van der Linden, W. J., Hambleton, R. K. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.

Verhelst, N. (2001). Testing the unidimensionality assumption of the Rasch model. *Methods of Psychological Research Online, 6*(3), 231-271. Institute for Science Education. Disponible de <http://www.mpr-online.de> (accesado el 20 de julio 2010)

Wilson, M. (2005). *Constructing measures: An item-response modeling approach.* Mahwah, New Jersey: Lawrence Erlbaum Associates.

Wright, B. D., Linacre, J. M. (1998). *WINSTEPS: A Rasch computer program.* Chicago, Illinois: MESA Press.

Zinbarg, R. E, Revelle, W., Yovel, I., Li,W. (2005). Cronbach's α, Revelle's β, and McDonald's ω: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1),123-133.

Anexo 1. Estadísticas descriptivas y correlaciones entre los aspectos medidos en EGRA en la versión de Nicaragua para el 2º, 3º y 4º grado y para la muestra completa

Grados		Aspectos medido	Mín	Max	Media	Desv. Est.	Coef. Var.	Asimetría	Curtosis	1	2	3	4	5	6	7	8
2º	1	Nombramiento de las letras	0	98.54	39.39	18.46	47	0.02	-0.20	1	0.29	0.22	0.76	0.72	0.71	0.64	0.60
	2	Identificación del sonido de la letra inicial de una palabra	0	10	3.99	3.21	80.52	0.17	-1.27		1	0.24	0.26	0.27	0.26	0.26	0.30
(n=2164)	3	Identificación de palabras que inician con el mismo sonido	0	10	3.56	2.24	63.02	0.18	0.14			1	0.22	0.18	0.23	0.18	0.21
	4	Lectura de palabras simples	0	108.23	34.77	20.59	59.22	0.16	-0.19				1	0.93	0.94	0.80	0.72
	5	Decodificación de palabras sin sentido	0	70.15	22.90	13.92	60.77	-0.05	-0.65					1	0.91	0.81	0.70
	6	Lectura de un pasaje	0	173.76	46.51	31.01	66.68	0.39	-0.11						1	0.82	0.73
	7	Comprensión de lectura de un pasaje	0	100	56.80	39.13	68.88	-0.46	-1.36							1	0.68
	8	Escritura de una oración	0	100	41.53	24.32	58.56	-0.07	-0.52								1
3º	1	Nombramiento de las letras	0	106.40	52.45	18.19	34.68	-0.08	-0.09	1	0.21	0.26	0.68	0.63	0.62	0.38	0.46
	2	Identificación del sonido de la letra inicial de una palabra	0	10	4.25	3.09	72.66	0.01	-1.20		1	0.30	0.19	0.18	0.21	0.19	0.24
(n=2218)	3	Identificación de palabras que inician con el mismo sonido	0	10	4.35	2.64	60.75	0.15	-0.43			1	0.25	0.22	0.26	0.14	0.27
	4	Lectura de palabras simples	0	139.87	55.05	21.26	38.62	0.14	0.42				1	0.87	0.88	0.45	0.55
	5	Decodificación de palabras sin sentido	0	88.88	35.09	13.50	38.46	-0.07	0.51					1	0.83	0.47	0.51
	6	Lectura de un pasaje	0	230.08	82.67	33.90	41.01	0.12	0.38						1	0.49	0.59
	7	Comprensión de lectura de un pasaje	0	100	81.80	24.87	30.40	-1.84	3.27							1	0.41
	8	Escritura de una oración	0	100	58.62	20.45	34.88	-0.29	0.17								1

Grados		Aspectos medido	Mín	Max	Media	Desv. Est.	Coef. Var.	Asimetría	Curtosis	1	2	3	4	5	6	7	8
4º	1	Nombramiento de las letras	0	172.67	62.53	17.56	28.09	0.05	0.77	1	0.15	0.25	0.61	0.57	0.53	0.19	0.31
	2	Identificación del sonido de la letra inicial de una palabra	0	10	4.41	2.99	67.96	-0.08	-1.12		1	0.29	0.14	0.12	0.13	0.09	0.20
(n=2267)	3	Identificación de palabras que inician con el mismo sonido	0	10	5.33	2.78	52.17	-0.14	-0.66			1	0.25	0.26	0.28	0.12	0.32
	4	Lectura de palabras simples	0	152.59	67.97	21.03	30.94	0.30	0.43				1	0.80	0.80	0.12	0.39
	5	Decodificación de palabras sin sentido	0	105	41.89	13.41	32.01	0.29	0.92					1	0.79	0.11	0.38
	6	Lectura de un pasaje	0	223.93	106.54	33.36	31.31	0.18	0.20						1	0.16	0.47
	7	Comprensión de lectura de un pasaje	0	100	86.94	17.93	20.63	-1.70	3.81							1	0.19
	8	Escritura de una oración	0	100	67.55	18.98	28.11	-0.33	0.18								1
Total	1	Nombramiento de las letras	0	172.67	51.64	20.39	39.49	-0.07	0.02	1	0.22	0.33	0.76	0.72	0.72	0.54	0.58
	2	Identificación del sonido de la letra inicial de una palabra	0	10	4.22	3.10	73.54	0.03	-1.21		1	0.28	0.19	0.20	0.19	0.20	0.25
(n=6649)	3	Identificación de palabras que inician con el mismo sonido	0	10	4.43	2.67	60.31	0.16	-0.44			1	0.34	0.32	0.36	0.23	0.34
	4	Lectura de palabras simples	0	152.59	52.85	25.01	47.32	0.09	0.04				1	0.90	0.91	0.61	0.66
	5	Decodificación de palabras sin sentido	0	105	33.44	15.70	46.95	-0.05	0.21					1	0.88	0.62	0.64
	6	Lectura de un pasaje	0	230.08	79.04	41.01	51.89	0.12	-0.25						1	0.63	0.69
	7	Comprensión de lectura de un pasaje	0	100	75.42	31.40	41.63	-1.40	0.90							1	0.59
	8	Escritura de una oración	0	100	56.10	23.89	42.58	-0.41	-0.08								1

Anexo 2. Estadísticas descriptivas y correlaciones entre los aspectos medidos en EGRA en la versión de Honduras para el 2º, 3º y 4º grado y para la muestra completa

Grados		Aspectos medido	Mín	Max	Media	Desv. Est.	Coef. Var.	Asimetría	Curtosis	1	2	3	4	5	6	7	8	9
2º (n=615)	1	Nombramiento de las letras	0	101.12	37.09	22.38	60.34	0.13	-0.73	1	0.43	0.26	0.76	0.74	0.67	0.64	0.30	0.59
	2	Identificación del sonido de la letra inicial de una palabra	0	45.71	8.53	8.63	101.14	1.09	1.23		1	0.25	0.40	0.38	0.35	0.33	0.13	0.23
	3	Recuerdo del sonido de las letras	0	10	1.04	2.16	208.12	2.21	4.17			1	0.29	0.29	0.30	0.27	0.19	0.16
	4	Lectura de palabras simples	0	100	24.84	21.65	87.18	0.43	-0.68				1	0.91	0.88	0.79	0.31	0.64
	5	Decodificación de palabras sin sentido	0	65.22	17.49	15.23	87.11	0.27	-1.08					1	0.84	0.79	0.31	0.65
	6	Lectura de un pasaje	0	190.19	36.16	33.82	93.54	0.74	0.33						1	0.81	0.28	0.63
	7	Comprensión de lectura de un pasaje	0	100	42.76	38.65	90.38	0.16	-1.54							1	0.39	0.60
	8	Compresión oral de un pasaje	0	100	39.54	30.86	78.03	0.26	-1.03								1	0.25
	9	Escritura de una oración	0	85	22.11	20.78	94.00	0.57	-0.68									1
3º (n=597)	1	Nombramiento de las letras	0	133.48	51.28	23.30	45.43	-0.07	-0.13	1	0.31	0.33	0.66	0.64	0.60	0.54	0.19	0.49
	2	Identificación del sonido de la letra inicial de una palabra	0	56.60	11.01	9.50	86.33	1.13	1.83		1	0.24	0.29	0.32	0.28	0.27	0.15	0.21
	3	Recuerdo del sonido de las letras	0	10	1.47	2.51	171.06	1.65	1.70			1	0.25	0.27	0.27	0.29	0.16	0.23
	4	Lectura de palabras simples	0	166.67	42.13	26.43	62.72	0.58	1.78				1	0.81	0.70	0.56	0.14	0.58
	5	Decodificación de palabras sin sentido	0	125.52	28.55	16.04	56.17	0.01	1.58					1	0.73	0.63	0.19	0.56
	6	Lectura de un pasaje	0	295.38	63.15	43.02	68.13	0.71	1.92						1	0.61	0.16	0.50
	7	Comprensión de lectura de un pasaje	0	100	63.79	35.91	56.30	-0.71	-0.88							1	0.26	0.42
	8	Compresión oral de un pasaje	0	100	53.07	30.00	56.54	-0.22	-0.93								1	0.18
	9	Escritura de una oración	0	95	33.26	20.80	62.54	0.16	-0.46									1

Grados		Aspectos medido	Mín	Max	Media	Desv. Est.	Coef. Var.	Asimetría	Curtosis	1	2	3	4	5	6	7	8	9
4º	1	Nombramiento de las letras	0	132.13	66.32	23.47	35.39	-0.40	0.73	1	0.11	0.18	0.47	0.57	0.46	0.31	0.16	0.29
(n=526)	2	Identificación del sonido de la letra inicial de una palabra	0	46.53	12.31	9.98	81.01	0.65	0.04		1	0.27	0.19	0.15	0.11	0.19	0.13	0.15
	3	Recuerdo del sonido de las letras	0	10	2.06	2.79	135.33	1.08	-0.06			1	0.16	0.26	0.18	0.20	0.24	0.22
	4	Lectura de palabras simples	0	196	60.36	30.65	50.78	0.67	2.08				1	0.60	0.55	0.32	0.14	0.43
	5	Decodificación de palabras sin sentido	0	136.36	37.94	15.82	41.69	0.48	4.45					1	0.59	0.36	0.15	0.36
	6	Lectura de un pasaje	0	295.38	96.73	45.75	47.30	0.35	0.83						1	0.43	0.14	0.46
	7	Comprensión de lectura de un pasaje	0	100	79.43	26.45	33.31	-1.57	2.02							1	0.33	0.28
	8	Compresión oral de un pasaje	0	100	60.42	28.32	46.88	-0.47	-0.67								1	0.11
	9	Escritura de una oración	0	100	46.60	21.08	45.23	0.11	-0.05									
Total	1	Nombramiento de las letras	0	133.48	50.81	25.87	50.92	-0.02	-0.37	1	0.32	0.30	0.70	0.73	0.67	0.60	0.31	0.55
(n=1738)	2	Identificación del sonido de la letra inicial de una palabra	0	56.60	10.53	9.48	90.03	0.96	0.98		1	0.27	0.32	0.33	0.28	0.31	0.17	0.22
	3	Recuerdo del sonido de las letras	0	10	1.50	2.52	168.33	1.60	1.47			1	0.27	0.31	0.29	0.29	0.23	0.24
	4	Lectura de palabras simples	0	196	41.53	29.92	72.05	0.73	1.51				1	0.81	0.76	0.63	0.29	0.62
	5	Decodificación de palabras sin sentido	0	136.36	27.48	17.74	64.57	0.20	1.01					1	0.78	0.69	0.32	0.61
	6	Lectura de un pasaje	0	295.38	63.76	47.65	74.74	0.65	0.69						1	0.68	0.30	0.61
	7	Comprensión de lectura de un pasaje	0	100	61.08	37.49	61.37	-0.59	-1.13							1	0.40	0.52
	8	Compresión oral de un pasaje	0	100	50.51	31.02	61.42	-0.14	-1.06								1	0.26
	9	Escritura de una oración	0	100	34.60	23.08	66.71	0.21	-0.51									1