FACULTAD DE CIENCIAS DE LA COMUNICACIÓN, TURISMO Y PSICOLOGÍA
ESCUELA PROFESIONAL DE PSICOLOGÍA
SECCIÓN DE POSGRADO

EVIDENCIAS PSICOMÉTRICAS PARA UNA BATERÍA DE HABILIDADES DE APRENDIZAJE PARA EL DESPISTAJE DEL RENDIMIENTO ACADÉMICO EN PRIMER GRADO DE PRIMARIA

PRESENTADA POR
CÉSAR AYAX MERINO SOTO

TESIS PARA OPTAR EL GRADO ACADÉMICO DE MAESTRO EN PSICOLOGÍA

LIMA – PERÚ

2014

Facultad de Ciencias de la Comunicación, Turismo y Psicología

Escuela Profesional de Psicología

Sección de Postgrado

EVIDENCIAS PSICOMÉTRICAS PARA UNA BATERÍA DE HABILIDADES DE APRENDIZAJE PARA EL DESPISTAJE DEL RENDIMIENTO ACADÉMICO EN PRIMER GRADO DE PRIMARIA

Tesis para optar el grado académico de Maestro en Psicología, mención Psicología Educativa

Presentada por:

CÉSAR AYAX MERINO SOTO

Lima, Perú

2014

Agradecimientos

En primer lugar, agradezco a mi padre, madre y hermana, por su siempre sobrevalorada confianza en lo que yo pudiera hacer.

A Marisol, porque fue el detalle que faltaba para completar una etapa de mi vida.

A mi hijo, Ayax, porque su existencia me lleva tratar de a ser un hombre mejor.

A ESSB y JMFA, por su consistente insistencia en tener esta meta cumplida.

Finalmente, a los sucesos (aparentemente) aleatorios que ayudaron a que las piezas del rompecabezas se unan (seguramente) en los momentos más oportunos.

Resumen

Actualmente, y sin riesgo de sobre-generalizar, en el Perú existen inadecuados instrumentos y procedimientos para efectuar evaluaciones a niños que ingresan al primer grado de primaria. El problema es crítico por las decisiones sobre los niños pueden ser equivocadas, además que la información derivada tendría pobre validez científica y social. La presente investigación tiene por objetivo obtener evidencias psicométricas de una Batería para el Despistaje de Habilidades de Aprendizaje para el Despistaje en Primer Grado (BDPG), orientado esencialmente al aprendizaje de la lectura y matemáticas, mediante la evaluación de habilidades de discriminación de letras y palabras, de elementos fonológicos y de estímulos visuales; conocimientos pre-matemáticos y comprensión verbal. Se obtuvo una muestra no probabilística multi-situacional en varios años consecutivos, y en dentro de cada año, participaron dos colegios públicos principalmente. Se examinó la validez de contenido mediante el análisis cuantitativo de ítems, la estructura y consistencia interna, la validez concurrente, de constructo y de diferenciación de grupos; las propiedades distribucionales (piso, techo y gradiente del ítem), el ajuste a una distribución psicométrica y las diferencias normativas. Los resultados fueron óptimos en cada uno de los análisis y subescalas, excepto para la escala de comprensión verbal, que obtuvo bajos coeficientes de validez interna y externa. Se discute las propiedades del instrumento de acuerdo a las expectativas teóricas y presupuestos estadísticos y psicométricos, así como los siguientes pasos en su investigación y uso.

Palabras clave: evaluación de despistaje, niños, lectura, matemáticas, validez.

Psychometric Evidences of a Battery for Screening of Learning Skills for the Academic Achievement in First Grade of Primary

Abstract

Currently, without risk of over-generalizing, in Peru there are inadequate instruments and scientific and conceptual framework for the screening assessment of children entering first grade. This problem is critical because the decisions on children can be wrong, in addition to the information derived has little validity. The aim of this research is to obtain evidence of a psychometric Battery for Learning Skills Screening in the First Degree (BDPG), essentially oriented to learning reading and math by assessing discrimination of letters and words, phonological skills, visual discrimination, pre-math skills and verbal comprehension. A multi-situational nonrandom sample was obtained in several consecutive years, and within each year was obtained, mainly involved two public schools. Content validity by quantitative item analysis, factorial structure, internal consistency reliability, concurrent validity, construct and group differentiation was examined; the distributional properties (floor, ceiling and gradient of item), the fit to a psychometric distribution and regulatory differences. The results were optimal in each subscale, except for the scale of verbal comprehension, which scored low coefficients of internal and external validity. The properties of the instrument according to theoretical expectations and statistical and psychometric assumptions, as well as the next steps in your research and use, are discussed.

Key words: screening assesment, children, reading, math, validity.

Tabla de Contenido

Introducción

En el capítulo uno se presenta primero la revisión de las investigaciones en que se evaluaron habilidades de aprendizaje de niños para el primer grado de primaria, encontrándose en ellas inadecuadas conclusiones e interpretaciones de los puntajes, además de problemas estadísticos y psicométricos en las investigaciones revisadas; estas investigaciones provienen especialmente de tesis de pregrado en psicología de varias universidades. Luego se describe la diferencia conceptual y práctica entre evaluaciones de despistaje y diagnósticas, en el marco de las evaluaciones de habilidades en niños que ingresan al primer grado. Se introduce aquí el impacto de habilidades más importantes que predicen mejor el rendimiento en lectura y matemáticas durante los primeros años de escolaridad; estas habilidades son la base del contenido de la prueba que se pretende validar (Batería de Despistaje para Primer Grado). Seguidamente, se presentan conceptualmente los procedimientos psicométricos importantes para el proceso de validación, limitándose a aquellos procedimientos aún poco conocidos y aplicados en el Perú, como el ajuste a una distribución psicométrica (beta binomial de cuatro parámetros), la verificación del piso y techo de una distribución, y la gradiente del ítem, entre otros. El capítulo uno cierra con el planteamiento del problema, que básicamente es la validación del instrumento en cuestión. El capítulo dos describe los procedimientos metodológicos realizados, fundamentalmente cuantitativos. Estos se sustentan en la teoría clásica de los test y en el modelamiento de ecuaciones estructurales; y cada análisis está directamente asociado a los objetivos del estudio declarados en el planteamiento del problema. El capítulo tres presenta los resultados ordenados de acuerdo a los objetivos. El capítulo final discute las evidencias psicométricas halladas, enfatizando la explicación de los resultados y su impacto sobre el uso del instrumento. La presente investigación cierra con declaraciones sobre las implicaciones para la práctica, limitaciones y recomendaciones.

CAPÍTULO 1
EVALUACIÓN DE NIÑOS PARA PRIMER GRADO: LA INVESTIGACIÓN EN PERÚ

CONTEXTO

En el contexto peruano, los estudios relacionados con el aprendizaje de niños preescolares y del primer grado de primaria, son generalmente tesis de pregrado no publicadas. Estos estudios se orientaron a la búsqueda de: a) correlatos cognitivos y concurrentes de las áreas de la madurez escolar, b) identificación de las características de aprendizaje de niños en niveles socioeconómicos específicos, y, c) la confirmación predictiva del aprendizaje de niños en el primer grado de primaria, enfocados primordialmente hacia la lectura. Tal interés también produjo evidencias de validez de constructo y datos descriptivos sobre los factores proximales y distales relacionados con el rendimiento del niño en el primer grado y en el nivel inicial (5 años). Por ejemplo, entre las habilidades muestreadas en estas investigaciones, se ha investigado la coordinación visomotora (Arraya, 1992; Piñasca, 1992), percepción visual (Espinoza & Matos, 1991; Llactahuacchaca, 1994), factores emocionales (Morote & Robles, 1978; Albiragorta, 1983) y la adaptación social (Chumbipuma, 1996). Descriptivamente, algunos estudios se focalizaron en el desempeño de niños del nivel socioeconómico bajo, y en las diferencias de género sobre el rendimiento (Majluf, Gutierrez, Delgado & Torres, 1971; Correa, 1977; Alzamora, 1981; Gutierrez, 1988; Vera, 1988; Sánchez, 1995) e inteligencia (Zensano, 1995). En estas investigaciones, no se encuentran explícitamente datos sobre validez de los instrumentos, confiabilidad de los puntajes, y el origen de las normas usadas para la clasificación descriptiva o la evaluación crítica de los resultados; por lo tanto, no se puede deducir la calidad métrica de los puntajes.

Se han encontrado sin embargo, algunas evidencias críticas sobre baterías de pruebas que tratan de predecir el rendimiento académico en el primer grado a partir de algunas evidencias halladas sobre factores cognitivos estudiados. Por ejemplo, Calderón (1991) cuestionó la validez predictiva de las pruebas de ingreso al primer grado, porque halló bajas correlaciones entre las notas escolares y las pruebas de ingreso administradas antes; Toledo

(1996) encontró correlaciones débiles entre el rendimiento escolar y los resultados de pruebas de habilidades específicas como inteligencia no verbal y vocabulario receptivo, administradas antes del inicio de la escolaridad; Quiroz (1976) también encontró bajas correlaciones entre los resultados de comprensión lectora y escritura, con los de una prueba específica cuyo objetivo era predecir estas habilidades específicas. Junto a estos estudios revisados, un producto atractivo para investigadores y profesionales es que algunos reportan normas den rendimiento de las pruebas utilizadas (Aranda, 1974; Canales, 1990; Espinoza, Piedra y Sotomarino, 1995; Rubio, 1992). Toda la información revisada en los párrafos anteriores no fue no ha sido publicada en revistas científicas.

Una breve información del contenido de las tesis revisadas se presenta en el Apéndice A. Sobre la base de éste se presentan los antecedentes del presente trabajo, que agrupa aspectos de carácter descriptivo y evaluativo de la investigación hecha sobre el tema en el Perú, los mismos que sirvieron para orientar los objetivos y construir hipótesis para la presente investigación, cuyo tema es el uso de medidas para estimar la preparación pre-académica de niños en el primer grado.

Transmisión de resultados

Un aspecto observado es el modo en que las investigaciones revisadas se conservan: casi todas han quedado almacenadas en las bibliotecas de sus universidades y no han sido publicadas en revistas científicas; por lo tanto, es evidente que estos estudios de tesis no han sido examinados por evaluadores externos, como usualmente se practica en las revistas científicas arbitradas. Debido a esta limitación no se puede tener seguridad de la validez de los procedimientos utilizados, los conocimientos derivados de ellos pueden tener sesgos y sus alcances son limitados en la intención de alimentar y robustecer la información científica en el tema. Estas afirmaciones se respaldan en las siguientes anotaciones metodológicas y psicométricas.

Problemas metodológicos

Se han observado algunos problemas metodológicos en las investigaciones revisadas que limitan la validez de sus resultados, por ejemplo:

Tamaño muestral. El tamaño muestral varió desde alrededor 50 participantes (Correa, 1977) hasta 288 (Chumbipuma, 1996). Este rango de tamaño muestral sugiere que muchos de los resultados estadísticos pueden carecer de poder estadístico, exhibir alto error de muestreo, ser inestables y carentes de precisión (Cohen, 2001). Aunque la estimación de la precisión a posteriori puede obtenerse mediante intervalos de confianza, el reporte de los resultados estadísticos descriptivos generalmente ha sido incompleto en las tesis revisadas, y por lo tanto su cálculo manual es imposible. También, no se exponía descripciones sustanciales de la muestra que podrían servir para conocer el grado en que la generalización de sus resultados podría ser efectiva.

Significancia práctica v. estadística. Los estudios revisados enfatizaron la significancia estadística como criterio para determinar la importancia de los resultados inferenciales, pero el enfoque de significancia práctica no fue declarado en ellos. Consecuentemente, los índices de magnitud del efecto tampoco estuvieron disponibles, o, cuando estaban presentes (por ejemplo, las correlaciones Pearson), no fueron cualitativamente descritas y contrastadas con la significancia de los resultados obtenidos de las pruebas de hipótesis estadísticas. Si se realizara el cálculo manual y preciso de los indicadores de significancia práctica, esto podría arrojar solo resultados aproximados, debido a que la información descriptiva es generalmente incompleta.

Control de variables. El control de variables moderadoras tampoco ha sido adecuadamente diseñado, ejecutado ni analizado estadísticamente. Por ejemplo, Calderón (1991) relacionó una batería de instrumentos de evaluación infantil (*WPPSI, Calificación Koppitz para el Dibujo de la Figura Humana*, el *Test Gestáltico de Bender* y el test *ABC*) con una prueba pedagógica ad hoc; pero no incluyó el puntaje de inteligencia como covariable entre las pruebas de madurez y

el rendimiento escolar. La inclusión y definición de variables potenciales para el control estadístico sirvió más para caracterizar a la muestra sobre las relaciones o diferencias exploradas. El control estadístico de variables irrelevantes en la determinación de la correlación aumenta la probabilidad que la correlación sea espuriamente alta (Chen & Popovich, 2002). Sin un control aproximado de la varianza irrelevante, entonces no se pueden extraer conclusiones válidas.

Confusión funcional entre tipos de evaluación. Las investigaciones revisadas no abordaron la diferencia entre evaluación de despistaje y evaluación diagnóstica (Woodburn & Boschini, 1995). La primera provee información del estatus de "riesgo". Los estudios revisados muestran comúnmente confusión sobre el uso de las pruebas, ya que los resultados directamente expresan el estatus de déficit o de capacidad en madurez para el aprendizaje de la lecto-escritura, pero en ningún momento se denominan a los niños con bajos puntajes como "en riesgo". Estas diferencias conceptuales son el fundamento para usar e interpretar las pruebas de madurez, ya que definen el alcance y las limitaciones que el usuario debe tener en cuenta (Hasbrouck, 1990). Considerando que las pruebas en edad preescolar son más inestables en sus resultados y han sido sujetas a críticas constantes sobre sus propiedades psicométricas (Alfonso & Flanagan, 1999), la confusión funcional entre estos tipos de evaluación aumenta el problema de las prácticas e interpretaciones inadecuadas en la evaluación e investigación del aprestamiento para el primer grado. Este problema no parece ser algo local, puesto que está presente también en otros países (Gracey, Azzara, & Reinherz, 1984)

Escalamiento de los puntajes. Los resultados de las pruebas de madurez en los estudios revisados, generalmente se analizaron como variables ordinales (por ejemplo, niveles alto, medio y bajo), presentando las descripciones cuantitativas y análisis mediante porcentajes (por ejemplo, Majluf, 1971; Correa, 1977; Zensano, 1995; Chumbipuma, 1996). Esta práctica reduce artificialmente la magnitud de las correlaciones y su poder estadístico, ya que las escalas

ordinales son menos descriptivas y contienen menos información que las escalas intervalo (Chen & Popovich, 2002). Por otro lado, las investigaciones que categorizaron ordinalmente los puntajes, aplicaron pruebas de hipótesis como la prueba de independencia χ^2, sin considerar que este estadístico no es sensible al ordenamiento de las categorías, por lo tanto, aumenta la probabilidad de error *Tipo II* (Pett, 1997).

Junto a los problemas metodológicos descritos anteriormente, la revisión de la literatura en nuestro medio ayudó a detectar algunos problemas psicométricos que a continuación se describen.

Aspectos psicométricos

Caracterización técnica de los criterios utilizados. Las correlaciones frecuentemente halladas entre criterios de rendimiento escolar y pruebas estandarizadas de aprestamiento (por ejemplo, Zarzosa, 2003; Toledo, 1996) son de baja magnitud. Parece plausible que esto se deba a la heterogeneidad de los criterios y objetivos evaluados por el profesor, así como su naturaleza mixta (cuantitativa y cualitativa) y los criterios de calificación de las pruebas construidas por los profesores, que en conjunto influyen para que las correlaciones de validez difícilmente alcancen valores moderados (Thorndike, 1989). Los posibles problemas de confiabilidad también son explicaciones razonables de las bajas correlaciones, más si se toma en cuenta la influencia del error de medición sobre la covariación entre variables (Feldt 6 Brennan, 1989; Nunnally & Bernstein, 1995; Thorndike, 1989). Por otro lado, hay que considerar los hallazgos de Zarzosa (2003), quien señala que la inestabilidad de las calificaciones de un semestre a otro, parecen contradecir la experiencia profesional en el área educativa, donde las calificaciones obtenidas tienden a ser más bien altamente multicolineales y poco variables. Esta afirmación, sin embargo, requiere respaldo empírico y por el momento, sólo tiene una base anecdótica del autor de la presente investigación.

Estimación de la confiabilidad. Exceptuando los trabajos normativos de Espinoza et al. (1995) y Rubio (1992) para la *Prueba de Funciones Básicas*, y de Canales (1990) para la *Prueba de Pre-Cálculo*, en ninguno de los estudios de tesis revisados se reporta algún indicador de la confiabilidad de los puntajes en las muestras de estudio. Únicamente se reporta la confiabilidad tomando en referencia a las investigaciones anteriores o a la información de los manuales. Sin esta información, se desconoce el grado de precisión psicométrica de los puntajes y de los resultados estadísticos ocasionados, aspectos afectados por el error de medición (Feldt & Brennan, 1989; Nunnally & Bernstein, 1995; Thorndike, 1989). Coeficientes de confiabilidad así como los de consistencia interna o de acuerdo de intercalificadores, sirven como información necesaria para entender las limitaciones de un estudio, tomando en cuenta que la confiabilidad pone un límite a la validez y a las correlaciones (Thorndike, 1989); sin esta información, los resultados de los estudios reportados no pueden ser evaluados respecto a su precisión.

Las correlaciones no corregidas por atenuación. Debido que no se obtienen estimaciones del error de medición, los estudios revisados tampoco efectuaron ajustes de las correlaciones entre las variables, para desatenuar el efecto del error de medición. Las correlaciones desatenuadas presentan la verdadera magnitud de la relación sin el error de medición intrínseca a los puntajes obtenidos (Thorndike, 1989), y sirven para mejorar la precisión de los resultados, al menos teóricamente. Aún en los estudios que reportaron la confiabilidad (Canales, 1990; Espinoza et al., 1995; Rubio, 1992), no se aplicaron este enfoque de desatenuación de las correlaciones.

Información normativa. Los estudios revisados no reportan el origen de las normas que sirvieron para clasificar a sus sujetos de estudio, o utilizaron la información normativa original publicada en la prueba. Por ejemplo, los trabajos de Majluf (Majluf et al., 1971; 1984) estimaron los niveles de aprestamiento y madurez intelectual fundándose en las normas originales de las

pruebas *ABC* (Filho, 1960), *School Readiness Survey* (Jordan & Massey, 1967) y la *Prueba del Dibujo de la Figura Humana* (Harris, 1963). Pero, algunos estudios normativos también se han desarrollado para adaptar pruebas existentes, como la *Prueba de Funciones Básicas* (Haeussler & Marchant, 2003) en las investigaciones independientes de Espinoza, Piedra y Sotomarino (1995) y Rubio (1992); y la *Prueba de Pre-Cálculo* (Milicic & Schmidt, 1980) en Canales (1990). Estas informaciones normativas tienen gran valor en la práctica profesional y en la investigación aplicada, pero son algo restringidos pues estos estudios normativos no fueron publicados en revistas científicas. Luego de estas normas y baremos construidos, no existieron posteriormente otras tesis que revisen o generen normas para las pruebas. Se debe señalar que cada estudio puede tener datos descriptivos de valor normativo, pero la revisión efectuada indica que tales datos son incompletos o de menor poder descriptivo, debido al muestreo no representativo y a la insuficiente información social y demográfica del mismo. Por ejemplo, Correa (1977) reporta medias pero sin información sobre la variabilidad; más aún, Chumbipuma (1996) solo usa porcentajes de la clasificación de la prueba 5 – 6 (B) fundándose en normas uruguayas, las mismas que fueron construidas aproximadamente en la década de los 60s. En ninguna de estas investigaciones hay estadísticos de distribución (simetría y curtosis), y dada la antigüedad de las investigaciones los resultados tienen valor histórico más que práctico.

Validez. Los estudios revisados, no propusieron *directamente* la evaluación de alguna faceta de la validez de los instrumentos usados como predictores o criterios. La acumulación de los resultados de estas investigaciones pueden servir, sin embargo, como respaldo y como límites de la validez de las mediciones hechas usando instrumentos de aprestamiento, pero esta utilidad tampoco fue descrita. Este marco para interpretar los resultados también es importante porque indica el valor predictivo de un instrumento, aspecto que concierne directamente a su validez; por ejemplo, fuera del contexto peruano, un estudio guatemalteco (Salazar, Amón y Ortiz, 1996) usó la prueba *ABC* (Filho, 1960) y encontró que no reprodujo la validez predictiva

reportada por el autor de este instrumento. Otro aspecto relevante al contenido de las pruebas usadas es que algunas de sus subescalas poseen poca cantidad de ítems, y, por lo tanto, hacen un inadecuado muestreo de los contenidos (Thorndike, 1989; Nunnally & Bernstein, 1995), hecho que los autores pasaron por alto para moderar sus interpretaciones. Aunque estos instrumentos son de despistaje, sus manuales proponen interpretar las áreas junto con el puntaje total, cuando se utilizan para describir las habilidades del niño. Este error del uso de las pruebas para el despistaje, no considera que el puntaje total es más apropiado comparado con sus subáreas, ya que éstas últimas son menos confiables y estables, (Thorndike, 1989). Finalmente, podemos afirmar que, en general, estos estudios abordan indirectamente evidencias de validez de constructo, pero no los explicitan entre sus objetivos. Esto los llevó a dejar de explorar las relaciones internas entre sus componentes (ítems y subescalas) como evidencias de la estructura interna de sus instrumentos (Nunnally & Bernstein, 1995).

Procedimientos de evaluación utilizados

Los métodos de evaluación en las investigaciones observadas estuvieron en un estrecho rango de variabilidad, ya que mayoritariamente se usaron instrumentos únicos como estimadores del complejo constructo de madurez escolar. Este método usa un solo instrumento, que generalmente es estandarizado (Mercer, 1991). Contrariamente, fue menos frecuente el uso de otras estrategias evaluativas, como las escalas estandarizadas respondidas por los profesores. En el contexto peruano y latinoamericano, el instrumento único parece ser el método de elección en la práctica profesional y de investigación; y menos frecuente es el uso de baterías de pruebas y casi no existen métodos estandarizados basados en la percepción de profesor.

Pruebas de aprestamiento. En este tipo de pruebas se incluyen las pruebas de madurez *tradicionales,* como *ABC* (Filho, 1960), la prueba *5 – 6, Forma B* (Gastelumendi, Isasmendi, Slovak & Semeleng, 1977) y el *School Readiness Survey* (Jordan & Massey, 1967), conocido en Perú como la *Prueba Jordan – Massey*. Estas obtuvieron más atención entre los

investigadores y en la práctica profesional en el período de los años 70s y 90s. En nuestro medio, las investigaciones con estos instrumentos han sido descriptivos (por ejemplo, Majluf et al., 1971; Correa, 1977; y Zensano, 1995). Posteriormente se presenta la *Evaluación de las Habilidades Básicas para el Aprendizaje* (EHBA: Eyzaquirre, 1996), que es una herramienta poco utilizada en las tesis de pre-grado; pero de acuerdo a la experiencia profesional del autor, el EHBA es una prueba cuya utilidad para asociarse al rendimiento académico es cuestionable. Por otro lado, en Latinoamérica, se tienen los trabajos de Woodburn y Boschini (1993, 1997) en Costa Rica, quienes adaptaron la *Prueba Gestáltica Evolutiva de Aprestamiento Escolar Antón Brenner* (Woodburn, Boschini, Zeledón & Fernández, 2002); en Argentina, una prueba diagnóstica de habilidades básicas en niños de zonas urbano-marginales (Furlan & Alderete, 2004); y en México, los trabajos de Guevara y colaboradores (Guevara & Macotela, 2000; Guevara, Hermosillo, García, Delgado & López, 2007). Una de las últimas investigaciones de adaptación en el contexto del inicio al primer grado fue la *Prueba Predictiva de Lectura* (Bravo, 1997), con la obtención de normas en Lima (Meléndez & Morocho, 2007).

Test de inteligencia. En los estudios asociados a las *habilidades* preparatorias o cognitivas que podrían impactar sobre el rendimiento escolar, la evaluación de la inteligencia sí estuvo presente. En las investigaciones revisadas, una de estas fue la *Prueba del Dibujo de un Hombre* (Harris, 196), usada por Majluf et al. (1971), y Toledo (1996). La elección de esta prueba posiblemente se deba a la facilidad de aplicación y obtención de la estimación del nivel intelectual, razonablemente suficiente para los propósitos de sus estudios. No se hallaron investigaciones en que usaron escalas de evaluación especializada e individualizada, como la Escala de Inteligencia de Weschler para Niños - Revisada (WISC-R) y el Stanford-Binet. Las potenciales limitaciones de este método de evaluación intelectual no fueron puestas en relevancia por las investigaciones citadas (por ejemplo, normas peruanas, replicación de la

confiabilidad intercalificador y validez de sus ítems en muestras peruanas) y es altamente probable que sus conclusiones sobre los niveles de inteligencia sean parcialmente inválidas.

Test perceptivo-motores. El *Test Gestáltico Visomotor de Bender* ha sido la prueba de habilidad visomotriz más popular en nuestro medio; y específicamente el Sistema Evolutivo de Koppitz (1984). Arraya (1992) usó esta prueba visomotora, mientras Llactahuacchaca (1994) usó el *Método de Evaluación de la Percepción Visual* (Frosting, 1980). Aunque estas pruebas, y particularmente el Sistema Evolutivo de Koppitz para el Test de Bender, son muy conocidas en la investigación y la práctica profesional de los psicólogos, en los estudios analizados no revelaron una especial atención por algunas de ellas.

Test de Lenguaje. No se hallaron tesis que describieran las características de lenguaje comprensivo o expresivo, o las usaran como variables de estudio. Tomando en cuenta que el lenguaje comprensivo es una de las variables que impacta sobre el rendimiento (Merino & Muñoz, 2007; Terrones, Solís, Canudas & Díaz, 1994)

La percepción del profesor. Desde hace décadas, varias revisiones están de acuerdo que la predicción del profesor con respecto a la lectura en primer grado y a problemas en el rendimiento académico, es tan bueno como el uso de alguna batería para evaluarlas directamente (Berdizewski & Milicic, 1974; Ilg, Ames Haines & Gillespie, 1978; Condemarín, 1989; Kenny, 1992; Kenny & Chekaluk, 1993). Pero en nuestra revisión encontramos que ningún estudio publicado ha utilizado alguna escala estandarizada que recoja la percepción del profesor sobre el rendimiento del niño. Hace varias décadas, la percepción del profesor se ha puesto en relieve como un aceptable predictor de los niveles alcanzados por los niños en pruebas estandarizadas de desarrollo, rendimiento y habilidades. Por ejemplo, Ilg et. al. (1978) encontraron concordancia entre las clasificaciones de los profesores y el Examen del Desarrollo elaborados por ellos en el Instituto Gesell; estas compatibilidades van desde la educación inicial (la correlación más alta) hasta el tercer grado incluso. Sin embargo, en las investigaciones de

tesis peruanas, no se consideraron como procedimientos para la recolección de datos. Hasta la fecha del presente documento, se está culminando la exploración psicométrica de una escala para profesores (Kenny & Izard, 1993; Kenny & Chekaluk, 1993), en el que se ha replicado la validez predictiva, la estructura interna y la consistencia interna en niños peruanos (Merino, 2008b).

Objetivos de evaluación

En el revisión de los procedimientos de evaluación en el Perú, una parte de estos se concentraron en la identificación de niños en riesgo de presentar problemas futuros en el aprendizaje de la lectura, situación que siempre tuvo los problemas de costo y tiempo para administrarlos (Salvesen & Undheim, 1994); por tal motivo, la creación de instrumentos de despistaje (screening) es un esfuerzo necesario. Si bien durante los años 60 y 70 en los EE.UU. se incrementó el interés en la identificación temprana de niños con incapacidad para la lectura, y por lo tanto la proliferación de instrumentos para estos fines, el interés por evaluar su efectividad no tuvo el mismo entusiasmo (Salvesen & Undheim, 1994). En el Perú, el interés y los problemas fueron paralelos; por ejemplo, en la década de los 70's aparecen los primeros reportes de investigaciones de tesis de pregrado, pero es el trabajo de Majluf et al. (1971) que fue el primero presentado en una exposición científica sobre la descripción de niños usando pruebas de madurez escolar; la evaluación administrada fue individual y usaron las que en esa época eran novedad: test *ABC* (Filho, 1960), *5 – 6* (Gastelumendi et al., 1977) y el *School Readiness Survey* (Jordan & Massey, 1967).

Alrededor de los 90s, se observa una sistemática orientación de la investigación hacia el mejoramiento de la tecnología evaluativa y un continuo interés por la adaptación y construcción de pruebas (Meza, 1988), el Instituto Nacional de Investigación y Desarrollo Educativo (INIDE) produjo los primeros trabajos elaborados en Perú para diseñar y validar materiales de evaluación de las nociones básicas del aprestamiento (Sánchez & Salazar, 1985)

y a las funciones intelectuales estrechamente asociadas a ellas (Palomino & Reyes, 1976). Estas medidas asemejaban pruebas de aptitudes y reflejaban la predominante conceptualización del desarrollo de las capacidades de los niños para ingresar al primer grado. Pero, estos intentos aparentemente no alcanzaron generalizarse en la práctica profesional ni en la investigación aplicada, ya que en las investigaciones revisadas en el presente estudio, publicadas o no publicadas, no los reportan.

A modo de conclusión, se puede afirmar que la investigación psicométrica sobre las medidas del aprestamiento de niños para el primer grado de primaria ha carecido de una continua atención científica, y por lo tanto, la distancia entre la investigación y la práctica profesional en el área se ha ampliado en los años. Los problemas metodológicos en el diseño y los análisis cuantitativos y psicométricos encontrados en los estudios, proporcionarían una cuestionable información para su posterior uso; este problema se ahondó por la elevada tasa de investigaciones no publicadas frente a las publicadas.

Adicionalmente, el surgimiento de instrumentos evaluativos en el área que nos interesa ha caminado lentamente, pero con aportes interesantes que tendían a definir constructos de amplio rango asumidos como descriptores y predictores apropiados de la lectura y escritura en el primer grado de primaria. Las investigaciones en nuestro medio no orientan sus hallazgos dentro de un marco de validez de constructo, y solo algunas pocas abordan directamente cuestiones de normalización y adaptación. El contexto descrito es un testimonio de las necesidades evaluativas para mejorar la medición y la práctica profesional en la evaluación de preescolares que ingresan al primer grado de primaria.

El siguiente paso en la prevención de problemas de orden académico es la planificación y creación de estrategias de identificación de niños con potencial riesgo de bajo rendimiento escolar. La evaluación que sirve para identificar tempranamente estos niños es un proceso en que el instrumento es un componente intermedio pero fundamental. El matrimonio entre la

ciencia y la práctica profesional se puede cristalizar con el desarrollo de un instrumento que cumpla con las exigencias psicométricas y que sea aceptado por los investigadores y profesionales. Por tal motivo, la presente investigación se dirige hacia la adaptación y evaluación psicométrica de una herramienta para la evaluación de niños que ingresan al primer grado, cuya finalidad es la detección temprana de problemas.

EVALUACIÓN DE DESPISTAJE Y HABILIDADES

En el plano internacional, las pruebas utilizadas en evaluaciones de la edad preescolar, entre 3 y 5 años de edad, han estado en constante crítica respecto a sus características técnicas que ponen los límites en la interpretación de sus resultados, y a su utilidad en el terreno profesional y de investigación. Bajo criterios cuantitativos y cualitativos desarrollados para evaluar su adecuabilidad en preescolares (Alfonso & Flanagan, 1999; Emmons & Alfonso, 2005; Alfonso & Flanagan, 2006), la evaluación de instrumentos cognitivos preescolares, ha arrojado resultados variablemente aceptables. Otras revisiones independientes han llegado a las mismas conclusiones. Por ejemplo, la revisiones hechas por Emmos y Alfonso (2005) sobre baterías preescolares de despistaje y la de Bracken (1987) sobre instrumentos preescolares de diagnosis individual, concluyeron que muchas de las pruebas revisadas poseen rangos variables de niveles psicométricos que van desde moderado hasta alto, respecto a la consistencia interna y estabilidad test-retest. Aunque las evidencias de validez provinieron de múltiples fuentes, los valores de validez también fueron relativamente inestables, ya que el tamaño de las muestras usadas en los manuales de las pruebas revisadas fueron pequeñas en muchos casos (Emmos & Alfonso, 2005). Otros aspectos que han sido criticados son la antigüedad de las normas antiguas, la diferencia de los contenidos de cada pruebas aunque la etiqueta nominal sea la misma-, la practicidad de su estructura, entre otros (Hasbrouck, 1990).

Evaluaciones de despistaje

Dentro de los usos destinados a las pruebas en la edad preescolar, el *despistaje* o *tamizaje* es quizás una de las funciones más importantes. El despitaje es un proceso de evaluación breve y de bajo costo, que busca detectar problemas e identificar sujetos en probable riesgo de una condición de interés (Hasbrouck, 1990; Meisels, Marsden, Wiske & Henderson, 1997). Es claro que la información generada de la evaluación de despistaje debe ser entendida de distinta manera que las pruebas diagnósticas, pues el tamizaje preliminar sólo aporta

indicadores para discriminar áreas con problemas, de un modo más general pero compendioso que otras pruebas más analíticas (Meisels et al., 1997; Woodburn & Boschini, 1995).

Mann y Powers (1997) tomando las recomendaciones de Meisels y Provence (1989), sostienen que la práctica de las evaluaciones despistaje hay que tomar en cuenta: a) El despistaje y la evaluación clínica deben ser vistos como parte de un proceso de intervención; b) Se debería incluir múltiples fuentes de información; c) El despistaje debe ser aplicado recurrente o periódicamente; d) El despistaje debe ser visto como un camino para una evaluación a profundidad posteriormente aplicada; e) Tanto los procedimientos de despistaje como de evaluación clínica deben ser confiables y válidos; f) Los miembros de la familia deben ser integrados al proceso de evaluación; g) El despistaje y la evaluación clínica deben ser aplicados en condiciones no amenazantes, naturalmente acondicionadas y con tareas relevantes al niño; y h) Todos los procedimientos implicados en la evaluación deben tomar en cuenta las variaciones culturales.

Otra diferenciación importante para la práctica profesional es respecto a las pruebas de desarrollo. Mientras que las pruebas de desarrollo están vinculadas con programas que enfatizan las habilidades básicas, las pruebas de aprestamiento o madurez escolar se relacionan con programas de enseñanza pre-académica (Hasbrouck, 1990), por lo tanto éstas serían más útiles en la programación de actividades recuperativas de conductos modificables por el aprendizaje.

Las pruebas de despistaje se diferencian de las de desarrollo en varios aspectos cualitativos y cuantitativos, referidos a la selección de sus ítems, su finalidad, su uso y aspectos psicométricos mínimos (Meisels et al., 1997; Meisels & Provence, 1989; Woodburn & Boschini, 1995). En Latinoamérica, algunos tests pueden referirse en esta modalidad de evaluación, especialmente en un estrecho rango de constructos, como los trastornos relacionados a las capacidades para aprender que evalúa el *Test de la Escuela Meeting Street* (Woodburn & Boschini, 1995) y el *Bracken School Readiness Assessment* (BSRA; Bracken,

2002); este último se orienta a niños desde los 2 hasta los 8 años, y evalúa conocimientos de conceptos básicos del niño referidos como parte esencial del aprestamiento. La *Prueba de Funciones Básicas* (Berdicewski & Milicic, 1976; Rubio, 1992) es otro ejemplo de instrumento para evaluar aspectos básicos y previos del aprendizaje general. En un rango más amplio de constructos, también existe el *Test de Desarrollo Psicomotor - TEPSI* (Haeussler & Marchant, 2003) para evaluar el desarrollo general. Las tareas seleccionadas para definir un instrumento de despistaje para el aprestamiento generalmente han consistido en obtener aquellos contenidos más predictivas para la lectura, matemáticas o escritura, pero con énfasis en la primera.

Aprestamiento y evaluaciones de despistaje

Entre los instrumentos que se enfocan en habilidades específicas fuertemente dependientes del desarrollo biológico, están las pruebas de integración visomotora, y entre las más populares se encuentran la Prueba de Integración Visomotora (Beery, 2000), y los sistemas de calificación adaptados para niños de la Prueba Gestáltica de Bender (Koppitz, 1984; Brannigan & Brunner, 2002) y particularmente para niños ingresantes al primer grado (Parsons y Weinberg, 1993; Sugar, 1995). Se debe hacer notar que en Perú, estas pruebas parecen ser usadas para tomar decisiones sobre el aprestamiento de los niños para el primer grado, lo es inválido de acuerdo a los objetivos de estos test.

Respecto a las pruebas de despistaje orientadas hacia la evaluación de la preparación del niño para iniciar la lectura presentan, su contenido y formato de las tareas son moderadamente similares, y ello hace difícil la comparación sustancial entre las pruebas (Hasbrouck, 1990). Estas diferencias pueden observarse no solo entre instrumentos distintos, sino al interior de cada una. Los diferentes formatos producen diferente dispersión de los puntajes, pues los ítems de una subescala pueden ser calificadas gradualmente. Los instrumentos más antiguos tienen esta característica problemática. Por ejemplo, una de estas ha provenido de Brasil, con la publicación de los test *ABC* (Filho, 1960), que es una de los primeras pruebas latinoamericanas publicadas

sobre la madurez escolar (Ardila, 2004); esta es sin duda una de las más resistentes a través de los tiempo, pues parece que las normas preparadas en esa época se continúan utilizando (al menos en Perú). Sin embargo, existen evidencias de su cuestionable valor predictivo (Ardila, 2004; Mora, 1999). La prueba de *Jordan - Massey* (Jordan & Massey, 1967) también ha probado ser resistente en el contexto latinoamericano, así como las pruebas *5 y 6* de origen uruguayo (Gastelumendi et al., 1977), lo que ha facilitado que en las tesis revisadas en la sección anterior, usen sus normas antiguas.

Otra generación de pruebas proviene de Chile con la adaptación (Abarca et al., 1965) del *Metropolitan Readiness Test* (Hildreth, Griffiths & McGauran, 1965), y posteriormente, con el desarrollo de la *Prueba de Funciones Básicas* (Berdicewski & Milicic, 1974, 2004). Más recientemente, se han desarrollado pruebas experimentales por parte del equipo de Bravo (1997, 2002, 2004) sobre habilidades fonológicas. Estas pruebas son comparativamente más homogéneas respecto al formato de presentación de sus ítems, lo que es un aspecto favorable en la reducción de la varianza debida al método de evaluación.

En Latinoamérica, no hay hasta la fecha una revisión cuantitativa ni cualitativa que realce las ventajas y limitaciones de estas pruebas, y que impactan en las prácticas de los psicólogos escolares y la validez y precisión de los resultados. Dada las limitaciones estas pruebas, se hace necesario no solo re-evaluar sus propiedades psicométricas, sino también la creación de instrumentos más sensibles, culturalmente relevantes, y que típicamente cubran las recomendaciones de las mejores prácticas sobre su elegibilidad y características cualitativas y psicométricas en pruebas de despistaje de habilidades pre-académicas (Hasbrouck, 1990) y de las pruebas que involucran habilidades cognitivas en general (Bracken, 1987; Alfonso & Flanagan, 1999; Emmons & Alfonso, 2005).

Madurez o habilidades

Aunque uno puede imaginarse una casi ilimitada cantidad de habilidades relacionadas, lo cierto es que tales áreas de habilidades y conocimientos que se creen predictores del éxito escolar, en un amplio rango de generalidad, incluyen el dominio de conceptos, procesamiento de información auditiva, visomotor y conciencia corporal; razonamiento y comprensión verbal, habilidades sociales y actividad motora y fina (Hasbrouck, 1990). Pero respecto al factor verbal, su mayor involucramiento parece ocurrir en edades más avanzadas, incluso desde el 2do grado, cuando las habilidades preceptúales son necesarias pero ya no suficientes (Solan, Mozlin, & Rumpf, 1985).

En un rango estrecho de habilidades, los predictores más conocidos para el aprendizaje de la lectura son la conciencia fonológica y el conocimiento de letras, memoria verbal, nombramiento rápido de objetos y conocimientos sintácticos-semánticos (Condemarín, 1989; Muter, 2000). Una revisión de 81 pruebas publicadas desde 1945 hasta 1990 diseñadas para evaluar las competencias del niño en aquellas áreas necesarias para aprender a leer y para decidir sobre la elegibilidad para el ingreso al nivel Kinder o primer grado, reveló que las áreas más citadas fueron la percepción visual, discriminación auditiva, identificación de letras, reconocimiento de palabras y vocabulario (Educational Testing Service, 1991). Estas áreas estarían asociadas para facilitar y expandir el dominio las habilidades sintácticas y semánticas para el aprendizaje de la lectura inicial. Pero según las investigaciones actuales, uno de los predictores más poderosos validado en varias meta-análisis internacionales es el desarrollo fonológico (Bravo, 2002), que es la habilidad principal que sirve para establecer un umbral lector (Bravo, 2002, 2004; Velarde, 2004) que ayude a diferenciar niños respecto a su capacidad temprana para la lectura.

Madurez para el aprendizaje. El término madurez se aplica de la palabra inglesa "school readiness", que se considera como ha capacidad del niño para lograr demandas de tipo cognitivas, sociales, físicas y emocionales en el desarrollo de la escolaridad (Thomas, 2001). Está altamente asociada con el estado y momento óptimo para la adquisición de aprendizajes (Ortiz, Becerra, Vega, Sierra & Cassiani, 2010), concepción asociada a la aparición de desarrollos críticos del niño en una edad determinada; por lo tanto, estos logros evolutivos críticos podrían servir como indicadores de la madurez para aprender (Hopkins, 2005). Estas capacidades deberían ser logradas o mostrar madurez antes de su ingreso al primer grado, y ajustarse a las expectativas institucionales o locales de lo que se considera madurez (Thomas, 2001). Estos factores son de amplio rango y de carácter más estable, y asociados a aspectos evolutivos del niño y culturales generales. Los aspectos cognitivos se pueden ver consensualmente en instrumentos que abordan un amplio rango de capacidades, como ocurre con las pruebas *ABC* (Filho, 1960), el "*Jordan- Massey*" (Jordan & Massey, 1967), el *Gesell School Readiness Test* (Ilg et al., 1978) y la *Prueba de Funciones Básicas* (Condemarín et al., & Milicic, 1981). Otros contenidos más específicos que se consideran en este tipo de pruebas son, por ejemplo, los conceptos básicos medidos por la prueba de Conceptos Básicos de Boehm – Tercera Edición (Boehm, 2000) o la Escala de Conceptos Básico de Bracken (Bracken, 1998), que se han utilizado como evaluaciones despistaje de la madurez escolar (Dumont & Willis, 2007).

Dada la importancia del aprendizaje de la capacidad lectura, la maduración ha sido el marco conceptual que predominó en la temprana literatura sobre las condiciones más importantes para el aprendizaje de la lectura. Esta influencia conceptual debió impactar en las estrategias de enseñanza y en la elegibilidad del niño para iniciar el aprendizaje escolar. Esto último se relaciona con la decisión de retrasar el inicio del primer grado o la repitencia escolar en el primer grado, bajo la idea que la maduración y su interacción con las experiencias de

aprendizaje, harán que el niño esté más maduro (Shepard, 1997). Su influencia también alcanzó a la creación de instrumentos de evaluación de las funciones básicas asociadas con la maduración del niño. Por ejemplo, el marco conceptual de los instrumentos que evalúan la madurez se refiere a los factores sinérgicos hacia el aprendizaje de la lectura, como aspectos biológicos, sociales, y algunos aprendizajes específicos y habilidades contextualmente adquiridas (Condemarín et al., 1981; Ilg, Ames, Haines y Gillespie, 1978).

Habilidades. Sin embargo, el enfoque de madurez para el aprendizaje de las habilidades pre-académicas fue cambiado hacia la perspectiva de la interacción persona-ambiente. Este enfoque consideró como habilidades aquellas que el niño requiere para un exitoso inicio del aprendizaje escolar (Guevara, Delgado, Hermosillo, García & López, 2007). Las habilidades asociadas predictivamente con el rendimiento en la lectura y matemáticas se han operacionalizado en métodos de evaluación basados en pruebas integrales u otras variantes.

Desde hace varios años se ha reportado una masiva evidencia de investigación que señala el gran poder predictivo del nombrado rápido, la conciencia fonológica y procesamiento ortográfico sobre el desempeño en la lectura, y su inclusión conjunta en una batería de despistaje permiten que hagan una contribución independiente a la predicción de problemas tempranos en la lectura (Badian, 1994, 2001). En un trabajo relativamente reciente de Badian (2001), el reconocimiento automático de palabras logró también asociarse a la fluidez y comprensión de lectura incluso hasta el 3er grado, y es un predictor de la lectura después de los primeros grados.

Ciertas tareas fonológicas son sensibles a la habilidad del niño para el reconocimiento fonológico en edades tempranas; por ejemplo, las tareas que hacen trabajar a niño sobre las sílabas en lugar que sobre los fonemas con más confiables en el nivel inicial; y las tareas de rima en la instrucción preescolar son predictivas del futuro rendimiento lector en los primeros grados (Badian, 2001). Además, las tareas de detección de rimas y segmentación silábica

(producir golpecillos según el número de sílabas en palabras dictadas han comprobado ser medidas predictivas en el nivel inicial 5 años (Badian, 2001).

Por otro lado, el procesamiento ortográfico se asocia al reconocimiento de letras y números incorrectamente orientados, como en las tareas de la prueba de *Inversiones Derecha-Izquierda* de Jordan (Jordan, 1980). Esta influye al procesamiento de letras y patrones de letras dentro de palabra y partes de palabras, haciendo una contribución independiente en los tempranos grados escolares respecto al procesamiento fonológico en la adquisición de la lectura (Badian, 1993; 1994). Posiblemente se asocia también a las experiencias tempranas con la lectura (Badian, 1994), lo que sugiere la influencia de la estimulación ambiental y el nivel socioeconómico (NSE) en su desarrollo. El papel del NSE también es relevante pues aportan como factores de riesgo que potencian el efecto de déficits específicos en los niños y explican una parte de la validez incremental de pruebas de madurez (Fowler, & Cross, 1986). Los efectos predictivos del NSE han sido corroborados al registrarse aspectos de la familia (nivel cultural, dificultades de aprendizaje) y del propio niño, como el nivel intelectual y la historia de enfermedades graves (Fowler, & Cross, 1986; Manterola, Avendaño, Cotroneo, Avendaño & Valenzuela, 1986)

Por otro lado, algunas tareas fonológicas son más probables de surjan como una sensibilidad fonológica general y estarían menos influenciadas por las experiencias tempranas de lectura del niño; estas tareas son el reconocimiento del número de silabas (tapping). Efectivamente, el reconocimiento del número de sílabas se correlaciona significativamente con la lectura en primer grado (Mann & Liberman, 1984), y ya que el reconocimiento de dos y tres sílabas que se acomodan a las capacidades de los niños preescolares (Mann & Liberman, 1984), y son ideales dentro de una batería de evaluación. Se ha hallado que las tareas fonológicas que usan la rima administradas a niños de hasta 3 años, contribuyen a predicir las habilidades tempranas de lectura (Mann & Liberman, 1984).

Finalmente, respecto a las tareas de emparejamiento visual influencias predominantemente sobre las tareas de habilidades ortográficas tempranas (Badian, 1993, 1994), que consiste en el procesamiento de símbolos escritos y secuencias de símbolos, independientemente de la habilidad para leerlos. El fracaso en este tipo de tareas influye también en la identificación de letras, pues en general, los déficits en la percepción visual caracterizan a los problemas de lectura en los grados iniciales (Badian, 1993, 1994). Procesamiento ortográfico se asocia al reconocimiento de letras y números incorrectamente orientados, como en las tareas de la prueba de Inversiones Derecha-Izquierda de Jordan (Jordan, 1980). Se refiere al procesamiento de letras y patrones de letras dentro de palabra y partes de palabras. Este hace una contribución independiente en los tempranos grados escolares respecto al procesamiento fonológico en la adquisición de la lectura (Badian, 1993; 1994), y posiblemente se asocia también a las experiencias tempranas con la lectura (Badian, 1994, 2001).

METODOLOGÍA PSICOMETRICA APLICADA A LA EVALUACION DE HABILIDADES PARA EL PRIMER GRADO

En esta sección se hará explicaran los componentes del análisis psicométrico involucrado en la construcción y/o adaptación relevantes al objetivo del presente estudio. Los aspectos de validez y confiabilidad son ampliamente tratados en la literatura metodológica de construcción de pruebas, así que no se presenta una descripción completa de los mismos, sino tangencial y con énfasis en algunos aspectos. Estos aspectos sobre procedimientos y fundamentos pueden ser novedosos para la actual práctica psicométrica en el Perú. Esta información influenciará directamente los análisis presentados en la sección Resultados del presente estudio. Los temas que se tratan a continuación se refieren al análisis de ítems, confiabilidad, distribución de puntajes,

Análisis de ítems

El análisis de ítem conducido por el examen tradicional de la dificultad y la discriminación, pueden ser evidencias de la validez de contenido (Rathvon, 2004) respecto a sus relaciones internas o externas con constructos relevantes. Estos últimos métodos son métodos cuantitativos que suplementan la información de los métodos cualitativos, y que generalmente ocurren antes de la recolección principal de datos para las evidencias de confiabilidad y validez posterior. También es posible obtener estimaciones de análisis de ítems desde muestras piloto más pequeñas, pero las interpretaciones basadas aquí deberían hacerse con precaución ya que el error de muestreo es mayor cuando el tamaño muestral de las muestras pilotos es pequeño (Cohen, 2001). En resumen, los procedimientos del análisis de ítems pueden converger en ayudar a tomar decisiones sobre las unidades mínimas de la prueba, y fortalecerán las cualidades de confiabilidad y validez, y las de techo, piso y gradiente del ítem. La aplicación de métodos cuantitativos en lugar de métodos basados en el juicio principalmente da evidencias

suficientes para juzgar la relación de los ítems con sus constructos y son congruentes con otros aspectos del análisis psicométrico, como la consistencia interna o análisis factorial. Aunque las evidencias que aportan únicamente son sobre las relaciones internas del instrumento, es información apropiada en la etapa inicial de adaptación del instrumento.

Confiabilidad

En el marco de la teoría clásica de los tests, la confiabilidad se refiere a la precisión del puntaje, respecto a infinitas aplicaciones del test a un sujeto (Cronbach, 1951; Nunnally & Bernstein, 1995). Frente a esta situación hipotética, en la práctica se obtienen diferentes estimaciones de la confiabilidad, de acuerdo a cómo el investigador conceptúa el error de medición (1 – precisión). Ya que el error de medición es un componente inherente del puntaje observado, éste debe calcularse en los puntajes obtenidos en cada muestra (Feldt & Brennan, 1989).

El coeficiente KR20 es un caso especial del coeficiente α (Cronbach, 1951), que se aplica a ítems de naturaleza dicotómica. Entre los usos de los coeficientes de consistencia interna está el *error estándar de medición*, que es la variación aleatoria en la unidad de medición del puntaje del test (Nunnally & Berstein, 1995); sin embargo, ésta es conceptualmente un promedio de la variación aleatoria, por lo que se puede preferir estimar el *error estándar de medición condicional* (EEMC; Feldt & Brennan, 1989), que informa de la variación del error en diferentes niveles del puntaje del test (Keats, 1957; Lord, 1955). En este caso, el procedimiento de Keats (1957) para estimar EEMC corrige la sobreestimación del procedimiento de Lord (1955b), y usa el error de medición derivados de los coeficientes de consistencia KR20 y KR21 (éste último se aplica a ítems con igual dificultad). Existe un ajuste para el KR20 cuando los ítems son legítimamente unidimensionales, pero observan diferente nivel de dificultad; esto fue

creado por Horst (1953), y es recomendado para la práctica profesional y de investigación (Charter, 1995).

Distribución de los puntajes

Techo. Es el máximo puntaje estandarizado obtenido en una prueba, asociado al máximo puntaje directo obtenido por la competencia del niño en el contenido evaluado por la prueba. Una prueba, según las recomendaciones actuales (Flanagan & Alfonso, 1995; Rathvon, 2004) tiene un techo apropiado si puede proporcionar puntajes exactos arriba de 2DE, y si el máximo puntaje posible es alcanzado por la muestra.

Cuando se observan inadecuados techos en puntajes emplazados debajo de 2DE, generalmente se asocian a inadecuados gradientes de ítem (Krasa, 2007), y ello sugiere una dependencia entre estos aspectos. Una prueba con un techo debajo de 2DE puede ocasionar una subestimación de las habilidades en examinados sobresalientes, y esto tiene una repercusión mayor en la identificación de niños con habilidades sobre el promedio. En una prueba cuyos puntajes se basan en los errores cometidos, en que los puntajes bajos indicarán un mejor desempeño en el atributo medido, el techo será un aspecto crítico para identificar a niños con bajos desempeños. Las pruebas difíciles para los examinados, en hay menos probabilidad de alcanzar el suficiente número de respuesta correctas para distribuir el desempeño alrededor y/o mayor a 2DE, generalmente subestimará el nivel de desempeño. En la Figura 1 aparece un sumario de este criterio.

Piso. Es el puntaje estándar más bajo posible cuando un examinado obtiene un puntaje directo de 1. Cuando una prueba tiene un piso adecuado, permitirá identificar niños con déficits o puntajes muy debajo de la media; contrariamente, el piso inadecuado origina una sobre-estimación de las habilidades del niño, y posiblemente la ocurrencia de falsos negativos. Es decir, un niño con un puntaje estandarizado (M = 100 y DE = 15) de

79 correspondiente a un puntaje directo de 1, tiene su habilidad falsamente representada, pues podría haber sido cualificado con PE aún más bajo, pero por efecto de un piso muy elevado en la distribución de puntajes su desempeño estatus puede ser la de un falso negativo. La falta de precisión en los extremos de la distribución de los puntajes Otro efecto de un piso inadecuado es que los puntajes y sus cambios asociados se pueden interpretar como lo que el niño no puede hacer en lugar de lo que puede realizar (Grigorenko y Sternberg, 1999). Operacionalmente, el puntaje directo de 1 debería derivar en un puntaje estandarizado mayor a 2 DE debajo de la media. Por ejemplo, en test con una media de 100 y DE 15, el puntaje directo de 1 debería emplazarse en 69 o menos. En la Figura 1 aparece un sumario de este criterio.

Gradiente del ítem. Este aspecto psicométrico concierne a la variación conjunta de la distribución de los puntajes a lo largo del completo rango de puntajes que se pueden obtener en una prueba, y se refiere al grado de cambio del puntaje estándar como una función del cambio en el puntaje directo (Rathvon, 2004). Para este análisis, específicamente se evalúa el cambio en unidades de un punto en el puntaje directo correspondiente al cambio en el puntaje directo. Operacionalmente, el cambio de una unidad en el puntaje directo debe producir un cambio de no más de 1/3 de 1 DE del puntaje estandarizado (Bracken, 2000; Rathvon, 2004). Una adecuada gradiente permite hacer distinciones más finas entre los examinados, y por lo tanto es sensible a pequeñas diferencias en el puntaje directo que se traducen también en cambios pequeños entre uno y otro puntaje estandarizado. Un largo incremento entre un puntaje estándar y otro no será sensible a estas diferencias a nivel fino. Esta característica se evalúa en todos los niveles de funcionamiento (Rathvon, 2004). Ya que los resultados de este análisis en los diferentes niveles de funcionamiento no podrían ser fácilmente comparados, se puede tener un marco de referencia común. El rango completo de puntuaciones puede dividirse

en segmentos de desviación estándar; esta estrategia fue intentada por Krasa (2007). Una extensión de este criterio aparece en la Figura 1.

	M (DE)	Techo ($X_1 \geq +2DE$)	Piso ($X_1 \leq -2DE$)	Gradiente	
				Nro unidades	Nro unidades entre piso y M
Puntaje t	50 (10)	70	30	3 puntos	≥ 8 puntos
Puntaje estandarizado 1	10 (3)	16	4	1 punto	≥ 6 puntos
Puntaje estandarizado 2	100 (15)	130	70	5 puntos	≥ 10 puntos

M: media. DE: desviación estándar. X_1: puntaje individual. Fuente: elaboración propia.

Figura 1

Criterios para cualificar el techo, piso y gradiente de una distribución de puntajes

Normas. Una de las estrategias para estandarizar los datos en obtención de interpretaciones normativas de los puntajes, es ajustar la distribución empírica de datos hacia una distribución teórica relevante. El ajuste realiza modificaciones en la forma de la distribución empírica, derivados de transformar los momentos de la distribución (como la media o los parámetros de la forma distribucional) para producir proporciones y puntajes ajustados a la distribución teórica (Angoff, 1971). Específicamente, el procedimiento transforma los puntajes observados usando como referencia las densidades de una distribución teórica hacia la cual se hace un ajuste; esta transformación es de tipo no lineal (Angoff, 1971). Uno de los principales efectos de estos ajustes es suavizar (smooth) las irregularidades de la distribución empírica (Kolen, 1991) y atenuar el efecto del error de muestreo.

Se pueden distinguir modelos paramétricos y no paramétricos para estimar las densidades de las distribuciones empíricas; las primeras especialmente se caracterizan por usar como referencia distribuciones teóricas con propiedades conocidas, como la curva normal (Yang, 2007); y cada una de ellos tiene parámetros fundamentales que se calculan de la información empírica para producir los cambios en la densidad de la distribución, como por ejemplo, la media, la dispersión, el valor mínimo o máximo, etc. Hay varios modelos estadísticos para hacer estos ajustes, como la distribución normal, pero hay otros modelos provenientes de conceptualizaciones más específicas en teoría de la medición como son los modelos psicométricos del grupo de las distribuciones binomiales. La aplicación de estos métodos asume que los puntajes son variables binomiales, y que producen una distribución discreta obtenido de respuestas correctas-incorrectas (Kolen, 1991). Uno de ellos es el *modelo fuerte de puntaje verdadero* (Lord, 1965), que declara que la forma distribucional del puntaje verdadero se expresa con un *modelo beta binomial compuesto de 4 parámetros* (4PB).

$$\Pr(X = i) = \int_0^1 \Pr(X = i \mid \tau) g(\tau) d\tau$$

La descripción extendida del método de estimación de los 4 parámetros basado en los momentos de la distribución empírica lo hizo Hanson (1991). El modelo 4PB ha probado ser una adecuada elección para el ajuste de puntajes de pruebas obtenidas de ítems dicotómicos (Lee, Hanson, & Brennan, 2002; Gempp & Chesta, 2007), y como generalización del modelo beta binomial, es más flexible y obtiene mejores ajustes frente a modelos no paramétricos (Cope & Kolen, 1990). La aplicación de cualquiera de estos modelos tiene como finalidad el mejoramiento de la estimación de las normas de los puntajes (Cope & Kolen, 1990), pero el proceso de aplicación de estos modelos se inicia

con la decisión sobre el más adecuado para hacer el ajuste. La elección de los modelos se hace con decisiones fundadas en el ajuste matemático documentado así como en la experiencia en el campo de aplicación de estos ajustes (Huber, 1999). Por otro lado, la inclusión de un parámetro significativo en la medición psicológica, hace que la elección del modelo 4PB sea más apropiado para poner en relevancia que los puntajes no son estimaciones libres de error de los atributos medidos. Este parámetro significativo es la confiabilidad.

PLANTEAMIENTO DEL PROBLEMA

La presente investigación se enfoca en las propiedades psicométricas de una batería de pruebas integrada para examinar las habilidades relacionadas con el inicio del aprendizaje de la lectura y escritura en el primer grado de educación primaria regular. Este instrumento es una adaptación de otra prueba de origen americano, que fue para el presente estudio fue sustancialmente modificada en sus contenidos, manteniendo el formato de presentación de los ítems en todas las áreas evaluadas.

La investigación ha sido motivada desde un horizonte práctico, respaldándose en a) la influencia de las habilidades específicas del niño que parecen predecir mejor el rendimiento lector y matemático en primer grado de primaria, b) la carencia de instrumentos actualizados y con buen fundamento científico para la detección de niños con riesgo de bajo rendimiento escolar, y, c) el estado de las investigaciones en el Perú sobre la evaluación psicológica para niños ingresantes a primer grado.

Considerando que un plan preventivo del bajo rendimiento escolar (necesario en nuestra realidad) debe incluir una evaluación de despistaje que responda a las exigencias de sensibilidad, especificidad y un balance costo-beneficio ventajoso de su uso, la propuesta de nuevos instrumentos es necesaria para desarrollar un buen trabajo profesional que esté de acuerdo con estándares científicos.

Actualmente, hay una extensa y diseminada preocupación por la buena práctica profesional en el uso de herramientas de evaluación psicológica, este hecho puede verificarse en las publicaciones que actualmente se consideran normativas para el desarrollo y conceptualización de instrumentos de evaluación, así como de sus mejores prácticas profesionales, por ejemplo, los "*Estándares*" de la American Educational Research Association, American Psychological Association, & National Council on

Measurement in Education (1999) y el documento del International Test Comission (2001).

La relevancia del presente estudio se fundamenta en dos aspectos; primero, en el matrimonio entre el soporte científico y el uso práctico inmediato de los resultados del presente trabajo, que se cristalizarán en un instrumento que se adecue con los estándares actuales en medición psicológica y educativa. En segundo lugar, el contexto de administración de las pruebas fue una situación típica del ejercicio profesional en el proceso de evaluación de niños, ésta no fue intrusiva o extraña a las actividades escolares de las instituciones participantes.

Formulación del problema

La formulación del problema de la presente investigación se presentada de la siguiente manera:

> *¿Cuáles son las evidencias psicométricas de validez y confiabilidad de un instrumento de Despistaje del bajo rendimiento académico en primer grado de primaria?*

El problema propuesto tiene varias facetas, que son los siguientes:

1. ¿Serán satisfactorias las evidencias de validez de contenido, estructura interna, relaciones convergentes, divergentes y concurrentes, y de diferenciación de grupos?
2. ¿La confiabilidad por consistencia interna será apropiada para los puntajes de las subescalas y puntaje total?
3. ¿Las características distribucionales de los puntajes (piso del puntaje y gradiente del ítems satisfacen a las mínimas recomendaciones técnicas?

4. ¿La distribución de los puntajes tendrá un satisfactorio ajuste a una distribución teórica psicométrica?
5. ¿Se observarán grandes diferencias en los puntajes del instrumento, entre los colegios colegio de procedencia de los participantes?

Objetivos de la investigación

Objetivo General

Demostrar las evidencias de validez y confiabilidad de la Batería de Despistaje para Primer Grado (BDPG) para la identificación temprana de niños en riesgo de bajo rendimiento académico al ingresar al primer grado de primaria.

Objetivos específicos

1. Proporcionar evidencias de validez de contenido mediante el análisis cuantitativo de la adecuabilidad psicométrica de los ítems (dificultad, discriminación y homogeneidad) de la Batería de Despistaje para Primer Grado (BDPG).
2. Proporcionar evidencias de la estructura interna la BDPG
3. Hallar evidencias de las relaciones divergentes y convergentes entre la BDPG y otros constructos,
4. Proporcionar las evidencias de validez de criterio de la BDPG, usando estrategias de relación concurrente.
5. Evaluar la capacidad de la BDPG para diferenciar de grupos de bajo y alto rendimiento escolar.
6. Estimar el grado de error de medición mediante la consistencia interna de la BDPG, así como su replicabilidad.

7. Identificar lo apropiado de las características distribucionales de los puntajes de la BDPG, respecto al piso y techo del puntaje, y de la gradiente del ítem.
8. Identificar el grado de ajuste de los puntajes con una distribución psicométrica.
9. Establecer el impacto de las diferencias demográficas (sexo del niño, estado civil, nivel educativo, composición familiar, centro de educación inicial de procedencia) sobre el rendimiento en la BDPG.

Algunas consecuencias teóricas y prácticas se pueden distinguir del presente estudio. Éstas se dirigen a la presentación de un nuevo instrumento contextualmente adaptado y sustentado científicamente dentro de los estándares actuales de investigación psicométrica. El punto de partida de la presente investigación es el vacío de instrumentos actuales, adaptados, predictivos y con buen soporte psicométrico para la evaluación de niños al ingreso del primer grado. La investigación podrá tener, sin embargo, repercusiones conceptuales sobre la evaluación psicológica escolar en el Perú. Las consecuencias teóricas confirmarán algunos hallazgos sobre la magnitud y comparación de los predictores del rendimiento escolar, como son las habilidades muestreadas por el nuevo instrumento presentado aquí; es decir, conocimiento de letras, discriminación fonológica, habilidad visomotora, conceptos cuantitativos y conceptos verbales.

Por otro lado, el presente estudio también explora las relaciones de validez de constructo entre las áreas evaluadas por el instrumento consideradas predictivas del rendimiento escolar y la intervención de variables moderadoras como la inteligencia, las conductas de aprendizaje y la habilidad visomotora. Por otro lado, las prácticas de evaluación en primer grado pueden ser influenciadas por el componente científico y riguroso que debe satisfacer un instrumento antes de utilizarse. La revisión de la literatura peruana (ver sección la sección del marco teórico), y las observaciones informales sobre

la práctica de las evaluaciones en primer por parte de los psicólogos escolares, ponen en relevancia la necesidad de presentar nuevas propuestas instrumentales y procedimentales. Dentro de un marco de prevención primaria, un instrumento parsimonioso como el validado en el presente estudio, podrá aplicarse masivamente a preescolares para identificar tempranamente el riesgo de presentar rendimientos escolares deficientes y problemas de tipo cognitivo, por ejemplo en niños las zonas marginales. Este diagnóstico inicial (screening) apoyará, por lo tanto, la identificación precoz de niños con alto riesgo de presentar futuros problemas de inadaptación y fracaso escolar.

CAPÍTULO 2
MÉTODO

TIPO DE INVESTIGACIÓN

La presente investigación consiste en la evaluación de las propiedades métricas de un instrumento de evaluación psicológica; por lo tanto, reúne procedimientos estadísticos descriptivos e inferenciales dentro del marco de la Teoría Clásica de los Test, y los unifica para evaluar las características tales como la validez y la confiabilidad, y aspectos más específicos como la gradiente del ítem, el piso de la distribución del puntaje y el funcionamiento diferencial del ítem. Los aspectos conceptuales y metodológicos para llevar estas estrategias han aparecido en publicaciones claves para guiar este proceso de validación (por ejemplo, Anastasi y Urbina, 1997; AERA, APA, NCME, 1999), de las prácticas de su buen uso (International Test Comisión, 2001), o guías de su evaluación (Bracken, 1987; Alfonso & Flanagan, 1999; Prieto & Muñiz, 2000; Lindley, Bartram & Kennedy, 2005; Flanagan & Alfonso, 2005). Aunque esta línea metodológica de presente investigación es precisa, se identifica dentro de las siguientes clasificaciones: una investigación transversal, con cohortes en diferentes puntos del tiempo.

De acuerdo al reciente esquema publicado por Montes y León (2007), el presente estudio se identifica como empírico e investigación instrumental. Este tipo de investigación observa las cualidades métricas de instrumentos de medición usando principalmente técnicas estadísticas. En general, se componen de estrategias correlacionales y de comparación para examinar los aspectos de validez y confiabilidad del instrumento en investigación. Tiene un componente de estudio cuasi-experimental, ya que en una de las comprobaciones de hipótesis del estudio de validez, se trata de establecer una relación causal entre las variables predictoras y un criterio de diferencia de grupos, pero no ocurre ninguna manipulación de variables independientes ni se aleatoriza la asignación de los niños a los grupos.

POBLACION Y MUESTRA

Descripción de la población de estudio.

La población son los niños y niñas en etapa de ingreso al primer grado de primaria de colegios estatales del distrito de Chorrillos. Estos niños han egresado recientemente de sus estudios preescolares en el nivel Inicial 5 años. La población estudiada pertenece a la jurisdicción del UGEL 07 (Unidad de Gestión Educativa), según esta registrada en el Ministerio de Educación. El distrito de Chorrillos es considerado como una población menos pobre (quintil 4), y sus indicadores de pobreza son menos severos que otros distritos (FONCODES, 2006); este distrito contiene áreas urbanas y peri-urbanas.

La población identificada puede ser considerada como una población finita, aunque de acceso difícil. La estimación de su número no es exacta debido a su variabilidad a través de los años. El registro del que se obtiene la estadística de matriculados tiene varias categorías para los colegios: Censo del Año Anterior, Datos Estimados y Omisos a la Estadística. Debe recalcarse que, durante el período de actividades académicas, ocurren numerosos ingresos, traslados o retiros de alumnos, sobre todo en los colegios no estatales. Además de esta advertencia, se debe considerar que: a) no todos los C.E. ingresan a tiempo en el directorio, antes de la edición, b) la edición del directorio demora su publicación, c) en esa moratoria de la publicación, en los C.E. pueden ocurrir retiros o nuevos ingresos; por lo que los informes recibidos por la USE pierden exactitud, y d) el directorio presenta un consolidado global, y, por otra parte, una síntesis de los alumnos matriculados en cada C.E., pero que no se subagrupan en los respectivos grados. Motivos por los cuales que no existe una estadística de alumnos por cada grado. Si agregamos a estos hechos la movilización de familias en términos de su residencia, la estimación que se logra es sólo aproximada respecto al potencial número de ingresantes desde las matrículas existentes en el nivel Inicial 5 años.

Descripción de la muestra de estudio.

La muestra de participantes provienen de cuatro instituciones educativas públicas de nivel primario (Colegios M. I., S. M., S. P.) e inicial (PRONOEI), ubicados en la zona urbana y peri-urbana de Chorrillos -distrito costero al sur de Lima Metropolitana-. Ambos colegios tienen enseñanza unidocente y los profesores son de género femenino. El tamaño de cada clase de primaria varía entre 30 y 35 alumnos, en el nivel inicial tienen 30 alumnos aproximadamente. Los alumnos que ingresan a estudios primarios en estos colegios no son seleccionados, ya que la matrícula es libre y universal (Merino, Díaz, Zapata & Benites, 2006). El sumario de las características demográficas y personales de los participantes, elementos clave para identificar a la muestra del presente estudio, se presentan en la

Tabla 1.

Las características funcionales, estructurales y organizacionales de los colegios de primaria pueden ser genéricamente similares, y por lo tanto se puede asumir que el monto y ritmo de experiencias de aprendizaje de los niños matriculados es similar.

No se pudo obtener más datos sociodemográficos de los participantes de los colegios de primaria, pero se hará una estimación cualitativa de otras características relevantes, considerando que el autor reside en la zona más de 30 años. En los colegios referidos, y considerando la zona donde se ubican, las familias de los niños generalmente alcanzan el nivel secundario; las madres tienden a pasar más horas con el niño que los padres, pues se ocupan del hogar y eventualmente realizan trabajos independientes.

Tabla 1
Descriptores y características demográficas de los participantes.

Descriptores	Características
Institución de procedencia	
Gestión	: Institución educativa pública
Enseñanza	: Unidocente
Tamaño de las clases	: Clases de primaria entre 25 y 30 niños
Zona de ubicación	: Urbana y peri-urbana, distrito de Chorrillos, zona sur de Lima.
Tipo de educación	Básica regular
Inclusiva	No
Actividades estimulativas programadas específicas	No
La muestra	
Género	: Niños y niñas
Nivel socioeconómico	: Predominantemente medio – bajo a bajo
Estructura familiar	: Predominantemente nuclear, cada niño viviendo con ambos padres
Actividad ocupacional de la madre	: Actividades del propio hogar y/o actividades económicas independientes
Actividad ocupacional del padre	: Trabajo independiente o empleado
Nivel de ingresos	: En promedio S/ 700 mensuales o menos
Edad	: Entre 5 años y 7 años
Estado civil de los padres	: Generalmente casados, pocos convivientes
Asistencia del niño a terapia	: Mayoritariamente, No
Residencia del niño	: Chorrillos, zona urbana y no semi-urbana

Mayoritariamente, son familias extendidas conformadas por los hermanos de los padres, abuelos, etc. Motivo por el cual los niños reciben actividades socializadoras de los padres y los demás familiares; estos hogares tienen más de tres miembros. Generalmente los padres son casados y pertenecen a la clase media baja o baja.

En relación de los niños del Programa No Escolarizado de Educación Preescolar (PRONOEI), este se ubica también en el distrito de Chorrillos. Los PRONOEI tienen la tarea de favorecer a los niños de familias económicamente pobres con aprendizajes pre-académicos mediante la estimulación de sus capacidades (Merino et al., 2006). Varios de estos programas están distribuidos en la zona urbana, pero principalmente en las zonas peri-urbanas del distrito; cada programa contiene varios locales en que realizan las clases preescolares, los participantes del estudio se ubican en la zona de la población peri-urbana, que es fundamentalmente de un nivel socioeconómico bajo o medio bajo.

Cada aula cuenta con una Promotora que posee estudios técnicos en educación, ella es quien organiza y conduce las clases conformadas por 10 a 20 niños aproximadamente. La gran mayoría de los niños y sus familias viven en los alrededores de los PRONOEI y han nacido en Lima, algunos de los padres son migrantes que provienen de la Sierra del Perú. El nivel educativo de las madres varía entre primaria y secundaria completa; el rango de edades de los padres al momento de la evaluación está entre los 26 a 30 años. La condición civil de la gran mayoría de las madres es la convivencia sin matrimonio oficializado; su ocupación principal se reparte entre las labores hogareñas y actividades económicas independientes. El ingreso económico se encuentra alrededor de los 700 soles mensuales o más, ya que la convivencia con miembros de la familia extensa aumenta los ingresos para la manutención básica. Generalmente, los niños y sus padres tienden a convivir con otros miembros de la familia materna o paterna, y por este motivo, los niños reciben actividades socializadoras de los padres y los demás familiares.

Como nota aclaratoria, el colegio S.P. se ubica en una de las áreas peri-urbanas de Chorrillos, específicamente en la zona llamada San Genaro. Es un colegio público con características similares en estructura y funcionamiento a los colegios públicos en áreas

como esta, así como el número de niños por aula. Las características demográficas de las familias de los niños muestreados aquí pueden ser consideradas similares a las descritas sobre los PRONEI, excepto que las condiciones económicas y de vivienda son relativamente mejores, ya que pueden superar los 500 soles mensuales como ingresos propios de la familia del niño.

Criterios de selección de la muestra, especificando el tipo de muestra.

La elección de estos colegios se hizo sobre la base de la disponibilidad y acceso del investigador, por lo tanto la muestra es de conveniencia y no probabilística; seleccionado por su disponibilidad y facilidad de acceso (Gall, Borg, & Gall, 1996). Usualmente, este tipo de muestra, como otras de naturaleza no probabilística, son elegidas cuando la naturaleza de la investigación es exploratoria (Gall et al., 1996); además, los resultados estadísticos y las conclusiones desprendidas de ellas sólo se pueden ver como una aproximación de los resultados en caso se aplicase un diseño propiamente probabilístico en el futuro (May, Masson, & Hunter, 1990). Además, la elección de una muestra de conveniencia tiende a ser razonable cuando se pretende describir las relaciones entre variables en un conjunto de condiciones específicas, sin que la meta sea necesariamente la generalización a la población (Jackson, 1999; Visser, Krosnink, & Lavrakas, 2000)

El valor científico de los estudios con muestreo no probabilístico no disminuye, una inmensa cantidad de investigaciones sociales y de laboratorio permiten describir fenómenos conductuales desde muestras accesibles, y luego, en estudios posteriores se evalúa su grado de generalización (Visser et al., 2000). Finalmente, los límites de generalización y representación poblacional de la presente muestra de participantes deben ser evaluados desde su descripción cuidadosa, pues es una invaluable información para

los usuarios, quienes decidirán el grado en que pueden utilizar el estudio para generalizarlo un contexto específico (Gall et al., 1996).

Criterios de inclusión y exclusión.

Los niños que fueron incluidos en la muestra debieron cumplir con algunos criterios de inclusión y de exclusión, así se tiene:

A. Los criterios de inclusión fueron:
    - Niños inscritos formalmente en el proceso de matrícula al primer grado.
    - Niños acompañados de sus padres o apoderados en el proceso de evaluación.
    - Niños que completaron todas las pruebas seleccionadas.

B. Los criterios de exclusión fueron:
    - Niños identificados con discapacidades auditivas, visuales o motoras severas.
    - Niños conductualmente indispuestos durante la evaluación.
    - Niños ingresantes para 2do grado o más.
    - Niños con las siguientes características detectadas durante el proceso de evaluación: elevada distraibilidad, severa habilidad para comprender las instrucciones,

Tamaño muestral

El tamaño muestral solo dependió del número de ingresantes, pero la valoración de lo apropiado de su tamaño provino de un análisis de poder respecto al mínimo número necesario para reducir el error Tipo II (Polit & Hungler, 1997). Este cálculo del tamaño muestral mínimo se hizo en relación de los estadísticos que se usarán para determinar los

valores de validez, como el coeficiente de correlación, la prueba *t de Student* para muestras independientes y la regresión lineal. Se aplicó el programa MINSIZE2 (Morse, 1999; Morse & Merino, en prensa) para estimar el tamaño muestral mínimo para detectar la significancia estadística en la correlación de Pearson, regresión múltiple y t de Student para muestras independientes.

Distribución de la muestra

La muestra total fue de 721 niños y estuvo compuesta por subgrupos de niños evaluados en años consecutivos desde 2004 hasta 2007. El tamaño muestral en cada condición (año, y/o colegio) varió entre 189 (año, 2004, ver

Tabla 2) y 40 (año 2007, ver Tabla 3). La variación del tamaño muestral en los años muestreados se debe al tamaño de la matrícula y a la afluencia de los niños durante el periodo de evaluación de ingreso al primer grado de educación primaria. Esto último significa que aún en los días del inicio de clases, un colegio (S.M.) extendía su matrícula a estos niños ingresantes, pero que no se beneficiaban de la evaluación y no están incluidos en la muestra de estudio. Mientras tanto, en el otro colegio muestreado (M.I), se cerraba la matrícula con el último niño evaluado. Los datos del año 2005 provinieron únicamente de un colegio (M. I.), mientras que en los demás años, se obtuvieron de ambos colegios (M. I. y S. M.).

A continuación, se analizaron las diferencias demográficas entre los periodos (*comparación entre-año*), y en cada colegio (*comparación intra-año*). Pero antes de iniciar la descripción de estos resultados muestrales, se debe indicar que en el año 2007 dos centros educativos también ingresaron a la muestra de estudio: un colegio estatal (identificado como S.P.) y un PRONOEI, ambos de zonas urbano-marginales. Los datos recogidos de los niños del PRONOEI y del colegio S. P. (Tabla 3) únicamente sirvieron para evaluar el desempeño de los niños y la observación de la profesora. Debido a limitaciones en el acceso a datos demográficos de los niños, solo se logró obtener la fecha de nacimiento proporcionado por los profesores. Ambas instituciones adicionan 71 niños a la muestra del año 2007, por lo tanto la muestra total en este año es de 238. Para informar de esta variación, en la

Tabla 2, los datos que aparecen dentro de los paréntesis contienen todos los colegios muestreados el año 2007 (es decir, M. I., S. M., S. P. y PRONOEI), y los números fuera del paréntesis, los correspondientes únicamente a los colegios M. I y S. M.

***Distribución entre-años*.** Analizando la distribución muestral entre los periodos (años) de evaluación, mediante el χ^2 de bondad de ajuste), la distribución de niños varones (χ^2 [3] = 4.05, $p > 0.05$) y de los que no asistieron a alguna terapia de aprendizaje (χ^2 [3] = 2.69, $p > 0.05$) fueron similares (

Tabla 2). Finalmente, respecto al orden de nacimiento, se compararon las distribuciones usando la prueba *T de Cramer –von Mises* (Abramson, 2004) con los siguientes resultados: la comparación de las distribuciones entre los años 2005 – 2006 (T = 0.42, p = 0.26) y, 2006 – 2007 (T = 0.26, p = 0.163) sugieren que no son estadísticamente diferentes. Las diferencias estadísticas se detectaron únicamente entre la distribución de los grupos 2005 y 2007 (T = 0.66, p = 0.015). Observando la

Tabla 2, estas diferencias provienen esencialmente de la primera celda (1er orden de nacimiento), mientras que el patrón general de frecuencias permanece similar.

Los datos presentados en la Tabla 3 son las distribuciones del género y la edad en el PRONOEI y el colegio S. P. recolectados en el año 2007. Comparando la proporción de varones (o de mujeres), estas solo muestran pequeñas diferencias con las proporciones observadas en la muestra total (

Tabla 2).

Tabla 2
Distribución de las características demográficas de los niños participantes, separados por año de recolección de datos[a].

	2004		2005		2006		2007[b]		Total	
	N	%	N	%	N	%	N	%	N	%
Sexo										
Varón	118	62.4	59	55.1	98	52.4	94 (132)	56.2 (55.4)	369 (407)	56.7 (56.4)
Mujer	71	37.5	48	44.8	89	47.5	73 (106)	43.7 (44.5)	281 (314)	43.2 (43.5)
Terapia del niño										
Sí	28	14.8	12	9.3	26	13.9	16	9.5	82	12.6
No	161	85.1	95	88.7	161	86.0	151	90.4	568	87.3
Orden de nacimiento										
1	--	--	35	32.7	89	47.6	100	59.9	224	48.6
2	--	--	35	32.7	54	28.9	39	23.4	128	27.8
3	--	--	17	15.8	27	14.4	18	10.8	62	13.4
4	--	--	6	5.6	10	5.3	3	1.8	19	4.1
5	--	--	4	3.7	3	1.6	1	.6	8	1.7
6	--	--	1	0.9	3	1.6	0	0	4	0.9
7 o más	--	--	0	0.0	1	.5	1	.6	2	0.4
Perdido	--	--	9	8.4	0	0.0	5	3.0	14	3.0
Total	189	100	107	100	187	100	167 (238)	100	650 (721)	

[a] : Ver la sección distribución de la muestra para explicación esta tabla.
[b] : La información entre paréntesis es el total de este año con los datos agregados de la Tabla 3.

Los datos de estos dos grupos se añadieron al cálculo de la muestra total en el presente estudio, y que aparecen entre paréntesis bajo el encabezado *2007*, en la

Tabla 2. Para propósitos de comparación rápida, los datos de la edad también aparecerán en la Tabla 5.

En resumen, la distribución del sexo, la asistencia a terapia, y el orden de nacimiento pueden considerarse concordantes entre los participantes de los periodos de evaluación muestreados.

Tabla 3
Distribución de participantes en dos centros educativos adicionales, en el año 2007

	PRONOEI (n = 31)		S. P. (n = 40)		Total (n = 71)	
	N	%	N	%	N	%
Sexo						
Varón	15	48.3	23	57.5	38	53.5
Mujer	16	51.6	17	42.5	33	46.4
	M	DE	M	DE	M	DE
Edad (meses)	74.3	3.5	77.0	4.0	75.2	3.7

Distribución intra-año. Para efectuar una descripción más minuciosa que pudiera revelar diferencias distribucionales entre los colegios, dentro de cada periodo de evaluación (comparaciones intra-año), en la Tabla 4 se presenta las características demográficas en los colegios M. I. y S. M. Para evaluar si existieron diferencias demográficas en cada colegio teniendo en cuenta el año, se examinó cada una independientemente.

Las potenciales diferencias entre proporciones se analizaron con la prueba χ^2 de Marascuilo (χ^2_{Mar}) para proporciones independientes (Marascuilo, 1970). Respecto al colegio M. I., éste contribuyó más a la proporción total de varones a través de los años (0.58); y entre los años, esta distribución de varones fue estadísticamente diferente (χ^2_{Mar} (3) = 9.035, p = 0.02), pero pequeña (V_{Cramer} = 0.15). Examinando el origen de esta diferencia por el procedimiento de Marascuilo (1970) para comparaciones pareadas entre proporciones independientes, esta solamente provino entre la muestra del año 2006 y 2007 ($p < 0.05$), pero igualmente esta diferencia proporcional fue relativamente pequeña (diferencia: 0.21).

Respecto al colegio S. M., se detectaron también diferencias estadísticas en la proporción de varones a través de los años (χ^2 [2] = 6.48, p = 0.03, pero sólo entre el

grupo del 2004 y 2007 ($p < 0.05$) en que la proporción de varones fue menor en este último año. El impacto de esta única diferencia fue, sin embargo, pequeño (V_{Cramer} = 0.12).

Finalmente, la información expuesta en la (Tabla 5) muestra que, entre los años, hubo diferencias estadísticamente significativas en la edad media (en meses) de los niños medido (ANOVA, $F[3, 717] = 972$, $p < 0.001$), pero estas diferencias fueron pequeñas ($\eta^2 = 0.039$). Se debe hacer notar también que dentro del periodo 2004 y 2006, los valores medios de variables en los colegios muestreados apenas se distinguen entre sí; sin embargo, las diferencias de edad sí fueron notorias en los niños de las instituciones PRO y S.P evaluados en el año 2007. Una vez más se debe aclarar que los datos que se pudieron comparar en los años anteriores al 2007 fueron los que se lograron recolectar la información demográfica en el momento de la aplicación de la batería de pruebas. Los colegios del año 2007 (S. P. y PRO), fueron recolectados luego del periodo de evaluación (entre los meses de marzo y mayo), y por este motivo las edades serán mayores que los niños evaluados en el periodo de ingreso.

En los años en que se pudo recoger la información del número de hermanos, éste varió claramente ($F[2, 529] = 52.12$, $p < 0.001$, $\eta^2 = 0.40$), es decir, del 2005 al 2007. Por otro lado, la edad media de los apoderados entrevistados en la evaluación (($F[3, 717] = 1.050$, $p > 0.05$, $\eta^2 = 0.004$), revelan magnitudes similares entre ellos en los años muestreados del 2004 al 2007. Por otro lado, el cálculo de la media y desviación estándar de la edad en meses, del número de hermanos y de la edad del apoderado entrevistado se realizó ponderando los valores total de cada año con el tamaño muestral correspondiente (Cohen, 2001).

Tabla 4
Distribución de las características demográficas de los niños participantes, separado por año de recolección de datos y los colegios muestreados

	2004				2005		2006				2007[a]				Total	
	MI		SM				MI		SM		MI		SM		N	%
	N	%	N	%	N	%	N	%	N	%	N	%	N	%		
Sexo																
Varón	60	62.7	58	62.4	59	55.1	46	48.9	52	55.9	56	70.0	38	43.7	369	56.8
Mujer	36	37.5	35	37.6	48	44.9	48	51.1	41	44.1	24	30.0	49	56.3	281	43.2
Terapia del niño																
Sí	13	13.5	15	16.1	12	11.2	17	18.1	9	9.7	8	10.0	8	9.2	82	12.6
No	83	83.6	78	83.9	95	88.8	77	81.9	84	90.3	72	90.0	79	90.8	568	87.4
Orden de nacimiento																
1	--	--	--	--	35	32.7	40	42.6	49	52.7	50	62.5	50	61.0	224	34.5[b]
2	--	--	--	--	35	32.7	35	37.2	19	20.4	16	20.0	23	28.0	128	19.7[b]
3	--	--	--	--	17	15.9	10	10.6	17	18.3	12	15.0	6	7.3	62	9.5[b]
4	--	--	--	--	6	5.6	6	6.4	4	4.3	1	1.2	2	2.4	19	2.9[b]
5	--	--	--	--	4	3.7	2	2.1	1	1.1	1	1.2	0	.0	8	1.2[b]
6	--	--	--	--	1	.9	1	1.1	2	2.2	--	--	0	0	4	0.6[b]
7 o más	--	--	--	--	0	0	0	.0	1	1.1	--	--	1	1.2	2	0.3[b]
Perdido	--	--	--	--	9	8.4	--	--	--	--	-	-	-	-		
Total	96	100	93	100.0	107	100.0	94	100.0	93	100.0	80	100	87	100		

[a]: No se incluye la muestra de los 71 niños descritos en la Tabla 3.
[b]: la suma de las frecuencias no sumará 650 que no está incluida la muestra del 2004, por falta de datos recolectados.

Tabla 5
Datos demográficos en variables continuas, de los niños y apoderados participantes

	2004			2005	2006			2007					Total	
	MI	SM	Total		MI	SM	Total	MI	SM	PRO	SP	Total		
Edad meses														
N	96	93	189	107	94	93	187	80	87	31	40	238	721	
Media	69.8	69.5	69.7	69.4	68.8	68.9	68.5	69.2	68.8	74.3	77.0	71.02	69.77	
DE	4.1	6.1	5.2	4.3	5.6	4.9	5.3	5.3	4.2	3.5	4.0	4.44	4.84	
Min	57	51	51	53	49	59	49	61	59	67	71	59	49	
Max	86	93	93	77	80	83	80	90	81	81	89	90	93	
Nro herm.														
Media	-	-		2.76	1.91	2	1.96	1.1	1.2	-	-	1.2	1.78	
DE	-	-		1.96	1.33	1.38	1.3	0.8	1.1	-	-	1.0	1.29	
Min	-	-		0	0	0	0	0	0	-	-	0	0	
Max	-	-		10	7	7	7	4	8	-	-	8	10	
Edad apod										-	-			
Media	34.8	32.7	33.7	34.1	34.3	34.5	34.4	32.2	33.9	-	-	33.08	33.80	
DE	8.3	8.2	8.3	7.3	8.4	9.2	8.8	6.4	8.1	-	-	7.28	7.72	
Min	22	15	15	15	18	19	18	17	21	-	-	17	15	
Max	66	57	66	54	76	67	76	57	70	-	-	70	76	
Nivel educativo	N (%)	N (%)		N (%)	N (%)	N (%)		N (%)	N (%)				N	%
Prim. Incomp.	0 (0.0)	5 (5.7)	5	3 (2.8)	2 (2.1)	2 (2.2)	4	2 (2.5)	4 (4.6)	-	-	6	18	
Prim. Comp.	3 (3.6)	5 (5.7)	8	3 (2.8)	4 (4.3)	6 (6.5)	10	1 (1.3)	2 (2.3)	-	-	3	24	
Sec. Incomp.	12 (14.3)	17 (19.6)	29	20 (18.7)	21 (22.3)	14 (15.1)	35	8 (10.0)	18 (20.7)	-	-	26	100	
Sec. Comp.	31 (36.9)	37 (42.5)	68	37 (34.6)	35 (37.2)	41 (44.1)	76	39 (48.8)	35 (40.2)	-	-	74	255	
Tec. Incomp.	4 (4.8)	9 (10.3)	13	14 (13.1)	12 (12.8)	16 (17.2)	28	7 (8.8)	7 (8.0)	-	-	14	69	
Tec. Comp.	23 (27.4)	12 13.8)	35	17 (15.9)	11 (11.7)	11 (11.8)	22	15 (18.8)	18 (20.7)	-	-	33	97	
Univ. Incomp.	4 (4.8)	0 (0.0)	4	6 (5.6)	6 (6.4)	1 (1.1)	7	3 (3.8)	3 (3.4)	-	-	6	23	
Univ. Comp.	7 (8.3)	2 (2.3)	9	7 (6.5)	3 (3.2)	2 (2.2)	5	5 (6.3)	0 (0.0)	-	-	5	26	
Perdido	12 (0.1)	6 (0.06)	18	0										

Luego de los análisis sobre los atributos demográficos de los participantes, se llega a siguiente descripción sumaria: la distribución del género, edad (en meses), asistencia a terapia y orden de nacimiento de los niños evaluados fueron esencialmente similares entre los niños muestreados en los cuatro años; una sección de la muestra, sin embargo, tuvo más edad dado que fueron evaluados posterior e independientemente a la muestra principal. Por otro lado, dentro de cada colegio, también hubo similaridad demográfica, y aunque se detectaron algunas diferencias estadísticas, estas fueron de pequeña magnitud como para producir diferencias en la distribución total. Respecto a las características de los apoderados, también existen similitudes en la edad y nivel educativo y generalmente fueron entrevistadas las madres de los niños. Se puede concluir que, exceptuando una pequeña sección de la muestra total, el resto puede ser demográficamente visto como un solo grupo dado que existen pequeñas diferencias entre ellas.

INSTRUMENTOS

Instrumentos para el proceso de validez de la bateria de despistaje

Los instrumentos aplicados se eligieron desde dos perspectivas: primero, para complementar la descripción de los niños en el proceso de evaluación profesional efectuada, se eligieron instrumentos de rápida aplicación y calificación y construidos para niños de edad temprana, pues el desarrollo tiende a variar aceleradamente durante el primer año escolar (Doig, 2005). Igualmente las evaluaciones se realizaron en períodos anteriores al inicio de clases, para evitar ser intrusivos e interferentes con las clases. En segundo lugar y desde una perspectiva de investigación, se seleccionaron varias medidas que aseguraban obtener puntajes confiables para el análisis de la validez. Los tipos de pruebas elegidos son 1) pruebas de desarrollo, 2) habilidades específicas, y 3) aptitudes e inteligencia. Seguidamente se describen cada una de ellas, y en la Figura 2 se presenta un resumen.

	Constructo			Criterio	
	Dif. Grupos	Convergente	Divergente	Predictiva	Concurrente
Habilidades Específicas	TPVNM Bender – SCC	Bender – SCC VMI			PGAB
Aptitudes e Inteligencia	DAP: IQ		PDP DAP: IQ TONI – 2		
Prueba de Desarrollo			TEPSI		

DAP-IQ: Prueba del Dibujo de una Persona para la Habilidad Intelectual. TPVNM: Test de Percepción Visual No Motor. Bender- SCC: Sistema de Calificación Cualitativa del Bender. VMI: Prueba de Integración Visomotora. PDP: Prueba de Aptitudes Preescolares. TONI-2: Test de Inteligencia No Verbal. TEPSI: Test de Desarrollo Psicomotor. PGAB: Prueba Gestáltica Anton-Brenner.

Figura 2
Resumen de la finalidad de las instrumentos criterio en la presente investigación

A. Validez de constructo:

A.1 Habilidades especificas

1. *Test de Desarrollo Psicomotor TEPSI* (Haeussler & Marchant, 1985).

Es un instrumento de despistaje del desarrollo psicomotor que evalúa, en forma gruesa, las áreas de Lenguaje (24 ítems), Coordinación Visomotriz (16 ítems) y Coordinación Motora Gruesa (12 ítems). El TEPSI consiste de materiales manipulativos y de lápiz-papel, además de un protocolo de registro. Se aplica individualmente a niños entre 2 y 5 años. Los ítems se califican con 1 ó 0. Se obtiene un puntaje para cada área, además de un puntaje Total. La consistencia interna reportada en el manual es mayor a 0.90 para el puntaje total. La validez se estableció con la varianza explicada por la edad, el sexo y nivel socioeconómico de los niños; es decir, los ítems están linealmente relacionados con la edad, los puntajes fueron diferentes para varones y mujeres, así como entre niños de diferente condición socioeconómica. Se han reportado investigaciones de validez de constructo (Pandolfi, Mathiesen & Oliva, 1997; Schonhaut, Maggiolo, Barbieri, Rojas, & Salgado, A. 2007; Schapira, Aspres, Benitez & Galindo, 1998); y en Perú, el TEPSI aparentemente tiende a ser el estándar en la detección de problemas del desarrollo psicomotor (Ministerio de Salud, 2005).

Originalmente se validó en Chile con 540 niños(as) de diferentes estratos socieconómicos de la región Metropolitana y de la Quinta Región. En el presente estudio, este instrumento se aplicará para respaldar la validez de constructo, específicamente la validez divergente.

2. Test de Percepción Visual no Motriz, TPVNM (Colarusso & Hammill, 1980).

Es una prueba estandarizada que evalúa en funcionamiento de la percepción visual, que fue diseñada en USA. El material se compone de una hoja de registro y un

cuadernillo de láminas. Los ítems se califican con 1 ó 0. El formato es de opción múltiple; cada ítem tiene 4 opciones de respuesta. La administración es individual, y toma aproximadamente 15 minutos o menos, aunque esto depende de la edad del niño. Se obtiene un puntaje total sumando los aciertos en las áreas de Figura-Fondo (8 ítems), Discriminación Visual (5 ítems), Memoria Visual (8 ítems), Cierre Visual (11ítems) y Relaciones Espaciales (4 ítems). La interpretación utiliza puntuaciones estandarizadas (Cociente Perceptual Visual) y edad perceptual. La confiabilidad por los métodos de estabilidad (entre 0.77 y 0.83) y consistencia interna (mediana = 0.80) fueron generalmente satisfactorios a través de las edades. En la validez de constructo, el manual reporta correlaciones con pruebas de inteligencia, de rendimiento, aptitudes y de visomotricidad en muestras de niños con diferentes cantidades (35 y 107) en cada grupo. La varianza compartida entre otras medidas de percepción visual fueron de mayor magnitud ($r_{mediana}$ = 0.49) que las demás pruebas ($r_{mediana}$ = 0.38 y 0.31 con pruebas de rendimiento e inteligencia, respectivamente), lo que corroboró la hipótesis del constructo medido. Estudios en el área de la rehabilitación ocupacional de adultos también dieron soporte a su capacidad para diferenciar personas adultas con déficits de percepción visual, teniendo lesiones cerebrales (York & Cermak, 1995; Mazer et al., 1998; Su et al., 2000).

La norma de estandarización original consistió en 883 niños normales entre 4 y 8 años y 11 meses, extraídos de 22 estados en los Estados Unidos de Norteamérica. Después de las normas originalmente publicadas en Estados Unidos, contemporáneamente se hizo un estudio con 183 niños peruanos entre 6 y 9 años de edad (Merino, Sotelo & Angulo, 2014), que reportó pequeñas diferencias en términos de magnitud del efecto y estadísticamente no significativas cuando se las comparó con la muestra de estandarización americana presentada en el manual. Una reciente revisión de la prueba reportó que, aún con su antigüedad, es una herramienta útil en la evaluación de niños

preescolares y escolares (Merino & Angulo, 2008). Este instrumento será utilizado para dar soporte a la validez de constructo, específicamente en la diferenciación por grupos.

3. Test Gestáltico de Bender – Modificado, SCC (Brannigan y Brunner, 2002).

Este instrumento evalúa el funcionamiento visomotor mediante el copiado de nueve figuras geométricas que se asumen libres de influencia cultural (Brannigan & Decker, 2003). La versión modificada contiene seis de los diseños originales (A, 1, 2, 4, 6 y 8) para su aplicación a niños preescolares hasta los primeros grados del nivel primario (4.5 hasta 8.5 años), dado que son los más apropiados para niños pequeños. El manual describe un sistema para puntuar el desempeño gráfico del niño, el *Sistema de Calificación Cualitativa, SCC* (Brannigan & Brunner, 2002) de 6 puntos, desde una puntuación de 0 (líneas aleatorias, garabateo, sin concepto del diseño) hasta 5 (representación exacta del diseño). Esta versión se califica por un método de inspección global, que refleja el grado de diferenciación y de la gestalt de los diseños reproducidos. El SCC acepta la administración individual o grupal, se hallaron diferencias pequeñas comparando los datos por el tipo de administración (Caskey & Larson, 1975; Brannigan & Brunner, 2002).

El manual presenta una extensa revisión de los hallazgos psicométricos, así como los criterios de calificación de cada diseño; por ejemplo la investigación sobre la confiabilidad (consistencia interna, test-retest e inter-calificadores) generalmente es elevada (≥ 0.80), y la validez del Sistema Cualitativo de Calificación da soporte a sus adecuadas propiedades métricas y sus cualidades instrumentales en la evaluación psicopedagógica, (Fuller & Vance, 1995; Brannigan & Brunner, 2002). Comparado con el Sistema Evolutivo de Calificación de Koppitz (1984), el SCC muestra correlaciones más elevadas con criterios de rendimiento escolar estandarizado (alrededor de 0.35), tanto

en los estudios americanos como en una muestra culturalmente diferente (en Hong Kong; Chan, 2002). Más recientemente, en el contexto latinoamericano se han efectuado estudios señalando que sus propiedades psicométricas originalmente satisfactorias son replicables desde estudios de confiabilidad inter-calificador (Merino & Benites, 2011) y de análisis no paramétrico de ítems (Merino, 2009); también, también se han reportado datos preliminares de validez de constructo como satisfactorias (Merino, 2010). El uso de esta prueba en la presente investigación respaldará la validez convergente (validez de constructo).

4. Prueba de Integración Visomotora, 4ta edición, VMI (Beery, 2000)

Es una prueba de evaluación de la habilidad para integrar el funcionamiento perceptual visual y de motricidad fina. Las tareas se presentan en un cuadernillo en el que están las figuras impresas, estas son consistentemente parecidas en su formato, ya que todas requieren que el niño copie figuras geométricas de dificultad creciente. Se califican dicotómicamente con 1 ò 0 para obtener un puntaje total. Hay dos formas, una larga de 27 ítems (aplicable a niños de 2 años hasta adultos) y corta (2 años hasta 7 años); en nuestro estudio se utilizó esta última. La validación de sus ítems y de su unidimensionalidad fue probada por métodos tradicionales y modernos (Teoría de Respuesta al Item); las estimaciones de consistencia interna a través de las edades son elevadas (> 0.85). Dado el extenso uso e información de su validez, es una de las pruebas más utilizadas para determinar el funcionamiento integrativo de la percepción visual y habilidad motriz fina (Kulp, 1999), especialmente por su relación con las habilidades pre-académicas en los primeros años de escolaridad (Kulp & Schmidt, 1996; Kulp, 1999). Su normalización se hizo en una muestra de 2614 participantes muestreados en USA. De

manera similar a la medida anterior, este instrumento ayudará a respaldar la validez convergente (validez de constructo).

A.2 Aptitudes e Inteligencia

5. Prueba de Diagnóstico Preescolar: PDP (De la Cruz, 1988).

Esta prueba elaborada en España mide aptitudes generales relevantes del aprendizaje preescolar. La PDP puede administrarse individual o colectivamente a partir el nivel preescolar (4 - 5 años), aunque el manual admite que la aplicación más óptima para el primer nivel (4 años) es a partir del último bimestre de instrucción preescolar. El tiempo de aplicación de la prueba completa es ente 45 y 55 minutos, administrada en dos sesiones y con una breve pausa. El material consiste en un cuadernillo en que todos los reactivos son pictóricos y exigen respuestas no verbales de trazado, delineado y rellenado. El PDP tiene 5 subpruebas (*Conceptos verbales* y *Conceptos cuantitativos, Coordinación Visomotora, Memoria auditiva y Aptitud perceptual visual*); la subprueba de Aptitud perceptual visual es la suma de los puntajes en *Constancia de la forma*, *Discriminación figura-fondo, Posiciones espaciales y Orientación espacial*. La calificación es objetiva y basada en ítems dicotómicos, y la puntuación se obtiene contando el número de aciertos en las tres primeras subpruebas; en otras pruebas, los errores son descontados de los aciertos obtenidos. El manual proporciona ejemplos y una descripción breve de los criterios para contabilizar o no un ítem, mediante una plantilla.

Recientes resultados publicados con este instrumento en niños peruanos indican moderados niveles de consistencia interna en la aplicación de grupo, para la muestra total ($\alpha_{prom.} = 0.72$), siendo similarmente consistentes entre varones y mujeres, pero de más baja confiabilidad entre niños de zona no urbana (vs. zona urbana) y de colegios públicos (Merino, 2008). Se ha demostrado sensibilidad para detectar diferencias de desempeño

aptitudinal entre niños de diversos niveles socioeconómicos bajo y medio (Merino & Muñoz, 2007); también se halló que los puntajes están levemente influenciados por la comprensión verbal en niños de bajo nivel socioeconómico (Merino, 2008a). Las normas exploratorias se obtuvieron de 320 niños desde un muestreo aleatorio de varios distritos del sur de Lima Metropolitana. En la presente investigación, este instrumento se orientará a dar evidencias de validez de constructo, particularmente de validez divergente.

6. Prueba de Dibujo de la Figura Humana para Estimación Intelectual en Niños, Adolescentes y Adultos: DAP: IQ (Reynolds & Hickman, 2004).

Esta es la más reciente prueba basada en el dibujo de una persona para estimar la habilidad intelectual aplicable en personas de 4 años hasta adultos (89 años). La instrucción para administrarlo consiste en solicitar al examinado que se dibuje a sí mismo; las instrucciones estandarizadas de calificación del DAP:IQ enfatizan los aspectos conceptuales del dibujo y no la calidad artística. El material consiste de un lápiz, hoja Bond A-4 y la hoja de registro. El sistema de calificación se basa en una puntuación cualitativa que varía desde 0 (parte no reconocible, incompleta o ausente) hasta 4 puntos (parte presente o cumpliendo todos los criterios anteriores); esta variación depende del ítem calificado. Los ítems forman 23 criterios estandarizados para calificar el dibujo: siete criterios son calificados con 0 y 1; siete criterios con 0, 1 y 2; ocho criterios con niveles de 0, 1, 2 y 3; finalmente, un criterio con niveles de puntuación desde el 0 hasta el 4. El puntaje se obtiene de la suma de las puntuaciones alcanzadas en los criterios; luego se deriva un puntaje de CI, que es una estimación cuantitativa y confiable de la habilidad cognitiva general. El manual reporta que los puntajes del DAP:IQ producen resultados estables ($r > 0.80$) en cortos periodos de tiempo (una semana) en una muestra de 45 personas de 6 a 57 años. Un reciente estudio en el ambiente Latinoamericano (Dávila,

Tello & Merino, 2009) han confirmado que la consistencia interna es replicable (≥ 0.79). Los estudios reportados en el manual mencionan que las correlaciones de validez concurrente con el Sistema Koppitz y Goodenough-Harris fueron 0.85 y 0.86, respectivamente; y con el WISC-III, r_{prom} = 0.51; y en el estudio latinoamericano, las correlaciones concurrentes con el Sistema Koppitz fueron elevadas, r_{prom} = 0.81 (Dávila et al., 2009); en este mismo estudio, las correlaciones con una prueba de razonamiento no verbal fueron más elevadas que con el Koppitz. La muestra normativa original del DAP: IQ fueron 3,090 personas de varias regiones de los Estados Unidos. Este instrumento se emplea para ayudar a establecer la validez divergente (validez de constructo).

7. Test de Inteligencia No Verbal: TONI – 2 (Brown, Sherbenou y Johnsen, 2000).

Es un instrumento estandarizado que evalúa la inteligencia no verbal mediante tareas de resolución de problemas con figuras abstractas como de emparejamiento, progresiones, analogías, clasificación, e intersecciones. El material consiste de una hoja de registro y dos cuadernillos de láminas, una para la Forma A y otra para la Forma B, pero sólo una se aplica. Ambas formas se consideran equivalentes. Su cualidad principal es que minimiza la influencia del lenguaje tanto para la respuesta del sujeto como en las instrucciones de administración, ya que se pueden usar mímicas para este propósito.

Se administra individualmente desde los 5 años hasta los 85 años, en un tiempo aproximado de 10 minutos o menos en niños de 6 años. Se califica dicotómicamente (1 ó 0) con ítems de formato de opción múltiple; se obtiene un puntaje total para estimar el nivel intelectual. En los datos originales americanos y en la adaptación psicométrica española, la consistencia interna fue alrededor de 0.90 para ambas formas, los ítems tendieron a ser más fáciles que las normas americanas originales. Las evidencias de validez respaldaron su peso no verbal frente a pruebas verbales de inteligencia verbal

(Brown, Sherbenou & Johnsen, 2000). Originario de USA, se estandarizó en más de 2000 personas, y en España, con más de 1000. Tiene dos formas equivalentes, cada una de 55 ítems. Un reciente estudio latinoamericano (Merino, Honores & García, 2008) en niños preescolares demostró que su consistencia interna es replicable, y que sus normas deben ser adaptadas debido a que sobre-estiman la estimación de la inteligencia. Similarmente a los anteriores instrumentos de aptitudes de inteligencia, ayudará a evidencias sus relaciones divergentes con la BDPG.

B. Validez de criterio

B.1 Habilidades especificas

8. Prueba Gestáltica de Antón Brenner, PGAB (Woodburn, Boschini, Zeledón y Fernández, 2004).

La PGAB es una prueba de evaluación cuantitativa y cualitativa de la disposición para entrar en la escuela, se fundamenta en los principios del desarrollo perceptual y conceptual del niño; su publicación original fue en 1967, en USA por Antón Brenner, el *Anton Brenner Developmental Gestalt Test of School Readines*. Las tareas son de breve aplicación y atrayentes para los niños, y requieren un manejo básico de números (Producción y Reconocimiento de números), copiado (Configuración Gestalt y Oración Gestalt) y el Dibujo de un Hombre.

El material consiste en un protocolo de registro, y hojas impresas para las tareas de copiado. Se aplica individualmente en un tiempo promedio de 5 minutos, a niños que ingresan a al nivel primario, entre cinco y seis años. Esta prueba es una versión hispana adaptada en Costa Rica por la educadora Sharon Wodburn desde 1987 (Woodburn, Boschini, Zeledón & Fernández, 2004). Los ítems se puntúan dicotómicamente con 2 o 0, y se obtiene solo un puntaje total. La confiabilidad por consistencia interna varió entre

0.91 y 0.71, siendo más confiable al inicio del primer semestre preescolar. Los estudios psicométricos sobre la validez han demostrado correlaciones positivas con la edad, y elevados coeficientes de confiabilidad intra y entre calificadores (≥ 0.70); la puntuación total correlacionó elevado con descriptores conductuales de problemas en el rendimiento académico hecho por profesores de primer grado (Woodburn et al., 2004; Zeledón, 1991). Asimismo, la validez predictiva en los niños de la muestra de normalización mostró que la clasificación correcta es mayor al 80%, absorbiendo elevada especificidad y menor sensibilidad (Woodburn et al., 2004). Correlacionalmente, las relaciones internas entre las tareas en general muestran ser moderadamente ortogonales. El estudio costarricense de normalización se hizo con 312 niños y para su interpretación se usaron niveles ordinales de desempeño. Esta prueba será importante para mostrar evidencias de validez concurrente (validez de criterio). Por lo anteriormente señalado, se considera que los instrumentos descritos buscan dar soporte a **la validez del instrumento** eje del presente estudio, es decir, la BDGP.

Instrumento de despistaje de rendimiento académico.

Batería de Despistaje para Primer Grado, BDPG (Merino, 2008a). Este instrumento explora habilidades pre-académicas para niños que están ingresando al primer grado de primaria; estas habilidades evaluadas exploran un estrecho rango de contenido sobre: Conocimiento de Letras y Palabras (18 ítems), Habilidades Fonológicas (17 ítems), Percepción Visual (21 ítems), Habilidades Cuantitativas (23 ítems) y Habilidades de Vocabulario y Conceptualización (20 ítems); para cada una de estas áreas se obtiene un puntaje, además de un puntaje total resultado de la suma de las subescalas. El BDPG se adaptó del *First Grade Screening Test* (FGST), Witheman (1987), en un proceso que se describe a continuación.

Versión original

La versión original del BDPG proviene del *First Grade Screening Test* (FGST) - PASS (*Program of Auxliary Services for Students*, Witheman, 1987), una prueba de 46 ítems creada para evaluar habilidades académicas en primer grado de primaria. El trabajo original no fue publicado para uso comercial, sino como un material interno de uso profesional en la ciudad de Philadelphia. Su calificación se hace contabilizando los errores en los ítems; es decir, se asigna 1 punto por ítem incorrectamente respondido, excepto en los ítems se seguimiento de instrucciones orales. Los ítems son de tipo opción múltiple, entre tres y cuatro opciones de respuesta. Exceptuando la tarea de Seguimiento de Instrucciones, todas las demás usan estímulos gráficos.

Respecto a la estructura interna del FGST, Whiteman hipotetizó una estructura tri-dimensional: Procesamiento Auditivo, Discriminación Visual y Concepto de Número/Letras. Los tipos de tareas incluidos en el FGST aparecen en la Figura 3. El instrumento se describió como una herramienta de despistaje, aplicable en grupos, de manera rápida y sencilla; de reducido costo en términos de tiempo y dinero para los materiales. El tiempo de aplicación de esta versión tomaba alrededor de 30 minutos (Whiteman, 1987). La elaboración de este instrumento fue guiado por fines prácticos, y en concordancia con la legislación educativa americana sobre la aplicación oportuna del proceso de evaluación que sirva para la identificación de niños con posibles necesidades de intervención educativa.

Dimensión	Tareas/Ítems
Procesamiento auditivo	• Discriminación de sonidos iniciales • Discriminación de rimas • Discriminación de sonidos finales • Comprensión auditiva • Seguimiento de instrucciones
Discriminación Visual	• Discriminación de figuras • Discriminación de palabras • Reconocimiento de patrones • Copiado de figuras geométricas
Concepto de número/Letras	• Conceptos de tamaño y cantidad • Igualar cantidades • Igualar números • Secuencia de números • Secuencia alfabética • Diferenciación de números y letras

Figura 3
Estructura y tareas correspondientes al *First Grade Screening Test* (Whiteman, 1987).

La obtención de los parámetros psicométricos y estadísticos del FGST se efectuó con datos recolectados durante tres años (1984 hasta 1986) en la ciudad de Philadelphia, en aproximadamente 45 colegios cada año y en una muestra de 972, 854 y 1158 niños (as), respectivamente. El estudio original de Witheman (1987) concluyó que el puntaje total es el resultado más confiable y replicable para la interpretación del rendimiento del niño, y no su hipotetizada estructura de tres dominios. En el análisis factorial que realizó, se usaron correlaciones Pearson, la regla de Kaiser para extraer el número de factores, y la rotación varimax para definir la estructura factorial. Se probaron varios modelos que variaron en el número de factores. Los resultados consistieron en ítems que cargaban en uno o varios factores, y los factores generalmente contenían ítems de diferente contenido.

Específicamente, su estudio factorial obtuvo una estructura de 17 conglomerados de ítems, que fue parcialmente consistente con el contenido y el formato de los ítems en el FGST. Por otro lado, la consistencia interna del puntaje total, basado en los ítems, fue > 0.78 en todas las muestras de recolección de datos.

Al evaluar la validez concurrente, Witheman halló que las correlaciones lineales con el nivel de lectura y matemáticas, así como la percepción de la profesora sobre el rendimiento del niño, fueron entre 0.40 y 0.50. La validez predictiva con medidas estandarizadas de rendimiento fue elevada (alrededor de 0.60) en un seguimiento de dos y tres años. Estos resultados sugieren que la evaluación de las habilidades de aprendizaje antes del ingreso al primer grado se relaciona efectivamente con el rendimiento académico, y puede servir para la detección temprana y derivación a intervención psicopedagógica.

Una observación crítica del reporte de Witheman sugiere que el análisis se hizo al nivel del ítem y mediante correlaciones Pearson, condiciones que son fuertemente influenciados por las características estadísticas de los ítems (Nunnaly & Bernstein, 1995). Este problema es conocido como "factores de dificultad del ítem", que surgen como consecuencia de la distribución de las respuestas. El reporte de Whiteman sobre su estudio factorial tiene, por lo tanto, problemas técnicos que condujeron a probar inadecuadamente la estructura hipotetizada.

Posteriormente, el FGST fue modificado ligeramente por Hirsh-Pasek, Hayson y Lescorla (1990) para utilizarse con niños de 5 años como una medida de criterio para ver sus correlatos con las expectativas parentales y de los profesores sobre el rendimiento preescolar. Esta versión no reportó estimaciones de confiabilidad y utilizó la recomendación de Whiteman (1987) sobre el uso del puntaje total.

Primera adaptación

La Figura 4 exhibe la estructura de la primera versión adaptada, que intentó capturar simétricamente las habilidades y estructura de la versión de Whiteman (1987). Esta versión respondió a necesidades prácticas de evaluación en el contexto profesional efectuada por el autor. El autor del presente estudio y una colega con experiencia en psicología eduativa inspeccionaron el cuadernillo original para hacer previsibles modificaciones relevantes al contexto peruano. De este modo, se decidió modificar los estímulos y varias de las instrucciones de las tareas, así como traducir las instrucciones del inglés al español. Una parte de los estímulos modificados correspondían a la discriminación de rimas, sonidos iniciales y sonidos finales, pues originalmente estos eran apropiados para el habla inglesa. La representación gráfica de las alternativas de respuesta, consecuentemente, también fueron modificadas para que la correspondencia entre la representación visual y la palabra-objetivos sea exacta. Otra parte modificada fue la discriminación e identificación de palabras, pues también se referían al habla inglesa. Finalmente, la tarea de diferenciación de números y letras fue modificada ampliando las opciones de respuesta correcta, para hacerlas homogéneas con el resto de los ítems. Se eliminó el subtest Seguimiento de Instrucciones porque no era apropiado para su aplicación grupal, además que se esperaba no estar conceptualmente relacionado con el contenido de las tareas del resto del instrumento.

Todas las modificaciones tomaron en cuenta la experiencia de ambos psicólogos, la opinión de profesoras de aula de 1er grado, la revisión de las pruebas existentes y la literatura científica sobre el tema. La elección de los estímulos que reemplazarían a los existentes fue consensuada respecto a la dificultad de respuesta y la facilidad de su representación, es decir seleccionándose aquellos con mayor probabilidad de ser respondidos (dificultad mayor o igual a 0.50) y requieran menor razonamiento en deducir

el significado de la tarea. Esta decisión fue consistente con los fines de la evaluación, pues se intentaba discriminar a niños con posibles déficits en los aprendizajes pre-académicos, aún con el riesgo de obtener un mayor porcentaje de falsos positivos.

Dimensión	Tareas		Ítems
Procesamiento auditivo	• Discriminación de sonidos iniciales • Discriminación de rimas • Discriminación de sonidos finales • Comprensión auditiva		2* 2* 2* 2
Discriminación Visual	• Discriminación de figuras • Discriminación de palabras • Identificación de palabras		4 4* 2*
Concepto de número	• Conceptos de tamaño y cantidad • Igualar cantidades (2 ítems) • Igualar cantidad a número • Secuencia de números • Secuencia de figuras • Diferenciación de números y letras		2 2 2 2 2 2*
Total	30 ítems		30

*: Nro de ítems modificados y adaptados en la primera adaptación

Figura 4
Estructura y tareas modificadas correspondientes al *First Grade Screening Test* (Whiteman, 1987).

Esta versión inicial fue probada en varios colegios, y se elaboró un estudio con 106 niños utilizando esta adaptación temprana, los resultados mostraron que un análisis de regresión lineal múltiple, predice alrededor del 0.35 unidades de las notas en primer grado sobre el rendimiento general, lenguaje y matemáticas, luego del primer trimestre de instrucción (Merino, en prensa). Sus correlaciones con una versión modificada de la

Prueba Gestáltica de Bender (Brannigan & Bruner, 2002) y con la Prueba de Factor G de Influencia Cultural Reducida (Cattell & Cattell, 1989; Altez, 1992) fue > 0.45. La consistencia interna KR-20 fue 0.81. Los ítems modificados y no modificados en esta versión sirvieron como criterio para hacer modificaciones a la siguiente versión.

Desde un punto de vista aplicado, esta primera adaptación fue bien recibida por los niños evaluados y satisfactoriamente aplicada en grupos de hasta 13 niños. La interpretación de sus puntajes se hizo normativamente, extrayendo la media y desviación estándar de la misma muestra del estudio. El único puntaje interpretable en esta versión fue la suma total de las respuestas correctas a los ítems.

Versión actual.

En la Figura 5 se describe la estructura tal como se presentó la BDPG en la administración del presente estudio.

Selección de ítems. La selección de los ítems y el formato fueron guiados por criterios conceptuales sobre los constructos derivados de la revisión de la literatura y de las pruebas existentes, así como por consideraciones metodológicas y prácticas racionalmente derivadas. Los criterios usados para seleccionar los ítems se describen a continuación:

Maximización de la varianza relevante al constructo. Los ítems elegidos deben contribuir a que el niño resuelva el problema planteado con conductas atribuidas al conocimiento y habilidades principalmente asociadas al contenido del área evaluada. La variabilidad de las respuestas ocasionada principalmente por las diferencias individuales en los conocimientos y habilidades del niño sería maximizada por ítems con características similares en el formato, tiempo de ejecución, etc.

Dimensión	Máx. puntaje	Tareas (numeración del ítem en la prueba)
Conocimiento de Letras y palabras	18	
Pág. 1: Primeros 12 ítems		
Combinaciones		(8, 9)
Letras		(10,11,12,13)
Sílabas		(14,15,16,17)
Escritura sílabas*		(18,19)
Pág. 4: Primeros 6 ítems		
Palabras		(48, 49, 50)
Escritura palabras*		(51, 52, 53.)
Habilidades Fonológicas	17	
Pág. 1: últimos 3 ítems		
Nro de sonidos		(20, 21, 22)
Pág. 2: Todos los ítems (14)		
Sonido inicial		(23, 24, 25)
Sonido largo – corto		(26, 27, 28)
Sonido final		(29, 30, 31)
Sonido final escrito		(32, 33, 34, 35, 36)
Percepción visual	21	
Pág. 3: Todos los ítems (11)		
Discriminación de Símbolos		(37,38,39,40,41)
Emparejamiento palabras		(42,43,44,45,46,47)
Pág. 4: últimos 10 ítems:		
Discriminación Igual – diferente		(54,55,56,57,58,59,60,61,62,63)
Habilidades Cuantitativas	24	
Pág. 5: Todos los ítems (10)		
Discriminación de números		(64, 65)
Conceptos cuantitativos		(66, 67, 68, 69)
Emparejar cantidad		(70, 71)
Simbolizar cantidad		(72, 73)
Pág. 6: Todos los ítems (14)		
Simbolizar cantidad		(74, 75)
Cardinalidad		(76, 77, 78, 79)
Secuencia		(80, 81)
Resol. Problemas		(82, 83, 84, 85, 86)
Habilidades Conceptuales: Vocabulario y Comprensión	20	
Pág. 7: Todos los ítems		(87, 88, 89, 90, 91, 92, 93, 94, 95, 96)
Pág. 8: Todos los ítem		(97, 98, 99, 100, 101, 102, 103 ,104, 105, 106)

* Forman un solo índice

Figura 5

Guía Resumida de ubicación, de las áreas y facetas evaluadas

Familiaridad. Algunos de los ítems requerían reconocer una figura significativa como medio para discriminar un aspecto asociado al problema planteado por las instrucciones. Por ejemplo, reconocer si una palabra inicia con

la letra “A” mediante la discriminación entre varias figuras, de una cuyo nombre se inicia con la letra “A”, en la instrucción se utilizó “avión”. La elección de las figuras debería estar incluida en la experiencia diaria del niño, expuesto a estímulos reales o figurativos, en medios audiovisuales o impresos. Este presupuesto fue validado parcialmente por la frecuencia de respuestas correctas.

Mantener la homogeneidad de respuesta básica y de calificación. Los ítems deberían calificarse de la misma manera, en un formato de opción múltiple, con solo una respuesta correcta y calificable como 1 para la respuesta correcta o 0 para la respuesta incorrecta, y que no involucre el juicio del examinador. La similitud en el formato de respuesta minimizaría la varianza irrelevante al constructo que podría ocasionar ítems de diferente presentación en el formato y las instrucciones.

Presentación sencilla. Los ítems propuestos deberían ser presentados con instrucciones fáciles de aplicar, no solo por psicólogos sino por profesores y asistentes con una preparación suficiente en la administración de pruebas estandarizadas. La presentación sencilla involucra el número de ítems por página, la cantidad de palabras en las instrucciones de aplicación al examinador y las instrucciones expresadas al niño, y, la cantidad de palabras presentadas para responder a cada ítem.

Instrucciones estandarizadas. Las instrucciones de aplicación para el BDPG serían estandarizadas y se modificarían luego de numerosos ensayos, y de las experiencias previas obtenidas. La elección y modificación de los ítems permitirían diseñar instrucciones de aplicación que fueron sencillas de usar, requerir el mínimo esfuerzo para su aprendizaje, y maximizar la varianza relacionada con el constructo.

Facilidad para la administración grupal. Los ítems seleccionados mantendrían las exigencias cognitivas y conductuales para ser respondidas principalmente en aplicaciones grupales. Aunque hay tareas que pueden ser más predictivas (como el nombrado rápido de figuras y la segmentación oral de fonemas), estas no se consideraron para su inclusión en el BDPG, ya que son efectivamente aplicadas en una evaluación individual. Ya que la finalidad principal del BDPG es la evaluación de despistaje, la aplicación grupal sería recomendada si los ítems pudieran aplicarse íntegramente en esta modalidad.

Muestreo de contenido. El número de ítems debería satisfacer la demanda de un apropiado muestreo de contenido y que favorezca la obtención de información confiable de las facetas medidas en ellas. Los ítems, capturarían las actividades que están vinculadas con el constructo que se está midiendo, en un nivel de dificultad esperado y vinculados con las demandas curriculares. El contenido elegido debería mostrar una indiscutible validez aparente o fascie, así como ser apropiadamente materializada en un formato que pueda ser aplicable grupalmente. Esta consideración afectaría el rango de contenidos que podrían ser incluidos en el BDPG. Aunque no se lograría un muestreo completo de todas las tareas, se elegirían las que puedan estar sustentadas teóricamente, de acuerdo a su semejanza con otras pruebas convergentes, y que requieran una respuesta sencilla por parte del niño (elección y marcado de respuesta).

Número de ítems. Este criterio fue moderado por cuestiones prácticas, como el tiempo de aplicación, el agotamiento potencial del niño, el costo de impresión y fotocopiado del cuadernillo, y por el aprendizaje del niño en la resolución de las tareas del instrumento. Por otro lado, se esperaba también que el análisis de ítems pueda indicar la remoción de algunos elementos, pero para evitar

una reducción sustancial de los ítems en cada faceta, el número de ítems fue maximizado considerando los criterios prácticos anteriores.

Efecto de techo. Debido que se procuraba construir ítems que tuvieran un rango de dificultad entre moderada y fácil, es posible que ocurriera un efecto de techo en la distribución de los puntajes. Por lo tanto, se creó un número suficiente de ítems que pudiera producir una efectiva variabilidad de los puntajes, especialmente en el nivel de pobre rendimiento en el instrumento, además de las consideraciones prácticas anteriores. Esto significaba introducir ítems de moderada dificultad.

Vinculación curricular. Los ítems deberían estar vinculados más directamente con el contenido del currículo educativo en el primer grado e iniciar 5 años, en la etapa de egresar de este nivel. Este vínculo permitiría probar directamente lo aprendido dentro de un marco de evidencias estandarizadas para reconocer los déficits de un niño. Este marco se fundamenta en las evaluaciones basadas curricularmente (currículum-based measurement), lo que hace al instrumento relevante al contexto peruano.

Cantidad balanceada en las áreas. Como la revisión bibliográfica ha mostrado, la evaluación de las habilidades pre-académicas se hace principalmente para las habilidades pre-lectoras. Sin embargo, la cantidad de ítems en el BDPG sería similarmente distribuida en todas las áreas evaluadas, valorándose la multidimensionalidad del constructo y el muestreo proporcional de su contenido entre las áreas.

Usos pretendidos. El BDPG fue construido apuntando a los siguientes cubrir los siguientes usos:

Proveer información para guiar las recomendaciones de intervención. En primer lugar, los contenidos seleccionados para el instrumento se relacionan con áreas relevantes para el aprendizaje de la lectura y matemáticas en los periodos iniciales de preparación para el primer grado de primaria. En segundo lugar, los puntajes contienen información estadística que compara el rendimiento del sujeto con un grupo normativo; y por último, los puntajes también se relacionan con criterios conceptuales y de rendimiento que permiten su interpretación significativa. Por lo tanto los niveles de rendimiento pueden orientar las sugerencias de intervención en las áreas de bajo desempeño, por ejemplo, los niños con puntajes debajo de una desviación estándar de la media serían identificados para requerir actividades correctivas.

Da información suplementaria y única para el propósito de despistaje. Debido al diseño del instrumento, éste se destina para detectar niños con potenciales problemas en el aprendizaje de la lectura y matemáticas durante el primer grado; además, ayuda a estimar las áreas que requieren atención diagnóstica más detallada y posibles orientaciones para la intervención recuperativa psicopedagógica.

Información normativa sobre las fuerzas y debilidades en las habilidades medidas. El instrumento contiene información diferenciada que permite discriminar varias habilidades y conocimientos que posee el niño, y que indican su desarrollo comparado con otros niños de edad similar. Por lo tanto, el rendimiento del niño comparado normativamente sirve para identificar el patrón de desempeño que puede impactar en los aprendizajes de lectura y matemáticas durante el primer grado.

PROCEDIMIENTO

Procedimiento de recolección de datos

Contexto general.

La recolección de datos se hizo en varias etapas en cada periodo de evaluación. Ningún periodo de evaluación fue la continuación del anterior, pues cada grupo de niños evaluados es un grupo independiente que requiere cumplir el proceso de evaluación para el ingreso al primer grado. Dentro de cada periodo de evaluación hubo dos etapas de recolección de datos, que fueron descritos en esta sección.

a) ***La evaluación de ingreso***. Ocurrió durante el proceso general de ingreso de niños para el primer grado de primaria en los colegios seleccionados. El servicio de evaluación de niños se solicitó al autor, como apoyo temporal durante el tiempo que dure el proceso evaluativo. El objetivo contractual de este proceso fue la evaluación de niños con el propósito de describir el patrón de habilidades pre-académicas cognitivas que afectan el futuro desempeño académico. Es un proceso que usualmente toma varios meses evaluar a todos los niños ingresantes, y la duración de la misma depende de varios aspectos, como la cantidad de vacantes, el número de niños inscritos para el proceso de evaluación, la organización del proceso, la extensión de los instrumentos de aplicación y la periodicidad de las evaluaciones. Los datos obtenidos desde los padres de los niños evaluados se hicieron paralelamente a la evaluación de los niños, de manera que mientras los niños eran evaluados en el espacio asignado, sus padres llenaban los cuestionarios seleccionados en otro ambiente, monitoreados por los asistentes de la evaluación.

***Formato de administración*.**

La evaluación se realizó en dos momentos: aplicación grupal y aplicación individual.

1. ***Administración grupal***. Para la aplicación de la batería de pruebas de grupo, se programó un máximo de niños en grupos de 8 ó 10 en todos los colegios participantes; sin embargo, el tamaño efectivo tendía a variar en el colegio M.I por dificultades organizativas de la institución, por lo que eventualmente el tamaño del grupo pudo llegar a 12 niños. En las fechas finales del proceso de evaluación, se alcanzó a tener grupos entre 3 y 6 niños. Generalmente, la aplicación del instrumento lo hacía el autor de la presente investigación, mientras que estudiantes de psicología apoyaban y monitoreaban al grupo. Luego de observar, participar y monitorear en las primeras semanas, en las siguientes los asistentes se encargaron de la administración completa. El espacio físico fue un aula vacía seleccionada previamente asignada por la Dirección del colegio. La obtención de la información desde los padres acompañantes de los niños se hizo en un aula separada, y dirigida generalmente por los asistentes. Esta modalidad tuvo las siguientes fases para aplicar la BDPG:
    a) *Presentación*. Se presentó las instrucciones generales de lo que los niños harían.
    b) *Identificar nombres en el cuadernillo*. Los niños que lograron escribir su nombre en la parte superior del cuadernillo, fueron identificados rápidamente por los evaluados. Cuando el niño no tenía esta capacidad, el examinador escribía el nombre y apellidos

luego de terminar la sesión de examen o durante el periodo de descanso intermedio.

c) *Ubicación de la primera tarea*. Se tuvo especial cuidado en la instrucción y asistencia de la ubicación de las primeras tareas a resolver. Ya que el formato de respuesta era similar en el resto de los ítems, los primeros ítems servían como entrenamiento.

d) *Monitorear la ubicación de la siguiente tarea*. La continuación exitosa de cada tarea suponía que el niño entendía el procedimiento. En caso contrario, se ayudaba al niño.

e) *Continuación de las tareas*. Se ayuda a los niños para pasar a la siguiente página.

f) *Revisión de respuestas erróneas*. La última fase, fue la revisión de los ítems sin respuestas, que consistía en aplicar individualmente el ítem no respondido. Esto aseguraba si la inicial ausencia de respuesta se debía a la distracción hacia otros estímulos, temporal falta de comprensión de la tarea o bajos conocimientos y habilidades para responder al ítem. Cuando en esta revisión el niño no respondía o lo hacía incorrectamente, amablemente se pasaba al siguiente ítem.

Las instrucciones de aplicación de BDPG aparecen en la ANEXO C.

Administración individual.

Se siguieron las instrucciones estandarizadas de los instrumentos de administración individual. Previamente, los asistentes tuvieron sesiones de entrenamiento y de estudio independiente del manual. Durante la administración, fueron monitoreados

por el autor durante las primeras semanas de evaluación, hasta que se juzgó que fueron competentes en el uso de los instrumentos.

Las características de los instrumentos respecto a la administración se describen en la Tabla 6.

a) Control de variables irrelevantes al constructo.

Se aplicaron algunos procedimientos para controlar variables irrelevantes a los constructos medidos, durante la administración de las pruebas. Ya que el proceso de administración de los instrumentos es un área considerada una fuente de potencial error de medición (Stanley, 1971; Lee, Reynolds & Willson, 2003)

1. ***Asistentes en la evaluación***. Se seleccionaron 4 asistentes de evaluación que apoyaron y, posteriormente administraron los instrumentos en las sesiones grupales y/o individuales. Cada asistente fue preparado en la situación descrita en el siguiente párrafo.
2. ***Conocimiento previo del material de evaluación***. Los asistentes de evaluación fueron capacitados en la estructura conceptual y funcional de los instrumentos, así como las reglas de administración y calificación. Los examinadores se familiarizaron con el contenido de la prueba. Para ello, los asistentes observaron al autor la aplicación *in situ* de la batería, en una o dos grupos de niños. En los siguientes de grupos de evaluación, los asistentes administraron la PDPG monitoreados por el autor; finalmente, los asistentes efectuaron la administración solos, sin presencia del autor. La batería es fácil de administrar, pero requiere múltiples tareas por parte del examinador: mantener la atención de los niños, no modificar las instrucciones estandarizadas, monitorear el avance del grupo, supervisar a

niños con posibles problemas de comprensión verbal, ansiedad o distraibilidad.

Tabla 6
Características específicas de la administración de los instrumentos

	Administración grupal	Administración individual
Pruebas	BDPG VMI DAP:IQ Bender – QSS Cuestionario demográfico (padres)	TONI – 2 PCAB
Evaluadores	Autor Asistentes	Autor Asistentes
Duración	40 – 50', incluiso 10 minutos de descanso.	15' – 25'
Pausa	Sí	No
Evaluados	Niños Padres	Niños
Momento	Primera etapa de evaluación	Segunda etapa de evaluación
Materiales	Proporcionados por los evaluadores, excepto lápiz, borrador y tajador	Proporcionados por los evaluadores, excepto lápiz, borrador y tajador

3. ***Instrucciones estandarizadas***. Las instrucciones para las tareas no se alteraron y se siguieron las alternativas de instrucción en cada tarea. Debido que la BGPG es un instrumento estandarizado, los procedimientos de aplicación y materiales deben ser las mismas para todos los niños y en

todas las condiciones. La aplicación no estandarizada introduciría, por lo tanto, una influencia no deseada sobre los puntajes.

4. ***El número de niños evaluados en grupo***: se determinó el número de niños por grupo evaluado fuera 10. El mínimo número fue 1, aunque este límite no pudo ser controlado en unos de los colegios, de tal modo que la variabilidad de niños evaluados en el colegio M.I fue mayor que la del colegio S.M.
5. ***Pausa en la evaluación***. Se programó una pausa de máximo 10 minutos en la administración de grupo. Esta pausa ocurrió inmediatamente después de aplicar la primera mitad del BDPG, que corresponde al último de la subescala de habilidades preceptúales ubicada en la 4ta página. Luego los 10 minutos, se pedía la asistencia de los niños y se los asignaba en los mismos lugares.
6. ***Número de sesiones***: aunque la batería completa puede administrarse en una sola sesión, es posible dividirla en dos sesiones según las características de los niños o las circunstancias el contexto. En nuestro estudio, exceptuando pocas oportunidades, la administración ocurrió en dos sesiones, especialmente cuando el grupo de niño era menos que 4.
7. ***Voz audible y regulada***. Los examinadores debieron enfatizar con el tono de su voz los aspectos de la instrucción que el niño usó para resolver la tarea. Por ejemplo, en la siguiente instrucción, la palabra subrayada deberá ser enfatizada:

   "Marquen con un <u>aspa</u> si estas dos palabras son <u>iguales</u>"
8. ***Contrabalanceo de las pruebas***. Las pruebas de administración grupal fueron parcialmente contrabalanceadas en el orden de aplicación; esta

aplicación parcial se hizo únicamente para las pruebas de covalidación, como el DAP:IQ y el VMI. Esta decisión se basó en que ambos instrumentos requieren el desempeño integrativo de habilidades motoras y visuales, y por lo tanto es plausible el efecto de entrenamiento de una prueba a otra ya que comparten varianza proveniente de estas habilidades en edades tempranas (Beery, 2000; Reynolds y Hickman, 2004). El contrabalanceo ocurrió durante las sesiones de administración grupal, se alternaron el orden de presentación de estas pruebas; después de estas pruebas gráficas, se administró seguidamente el BDPG.

9. ***Re-aplicación de ítems no respondidos***. Los ítems nos respondidos durante la aplicación del BDPG, se administraron nuevamente inmediatamente después de terminar la aplicación grupal. Esto se hizo para asegurar que los ítems sin respuesta no se produjeron por la falta de atención temporal, déficit en la comprensión, por aspectos estilísticos de respuesta del niño o por falta de conocimiento de lo que pedía el ítem.
10. ***Disposición espacial de los niños***. Los niños fueron ubicados separadamente en las mesas, de tal modo que la posibilidad de copia se minimizó. Cuando no hubo suficiente espacio, se colocaron entre los niños materiales (sillas) para no permitir la visibilidad del cuadernillo de los niños adyacentes.
11. ***Mantención la atención de los niños***. La continuidad de las instrucciones, y la capacidad de los examinadores para atraer la atención durante la administración se hizo efectiva durante todo el proceso de evaluación grupal. Adicionalmente, los asistentes monitorearon la evaluación para capturar también la atención del niño hacia la tarea e instrucciones.

b) Materiales

Para el niño: Se solicitó a los padres de familia que su niño tuviera dos lápices, borrador, tajador. El examinador proporcionaría los cuadernillos de la prueba y hojas Bond A-4.

Para el examinador: Cuadernillo de las pruebas, cuestionarios para los padres de familia, pizarra, lapiceros, Hojas Bond A-4.

Para el padre de familia: lapicero o lápiz.

Procedimiento para el análisis de datos

Dado que la investigación es psicométrica, se diseñaron varias estrategias estadísticas para la obtención de evidencias de confiabilidad y validez. Las estrategias tuvieron como base las recomendaciones técnicas de los estándares de medición psicológica y educativa (American Educational Research Association et al., 1999) y de la Comisión Internacional de Test (2001). Ambos documentos dan guías y recomendaciones para las prácticas apropiadas de reporte sobre las evidencias de validez y confiabilidad de los puntajes de instrumentos en construcción y/o adaptación. El enfoque de análisis psicométrico fue el de la Teoría Clásica de los Test (TCT). Este modelo asume que el puntaje observado contiene dos componentes aditivos: el puntaje verdadero y del error aleatorio. La ecuación conceptual y principal es la siguiente:

$$\mathbf{X = V + e}$$

X = Puntaje observado
V = puntaje verdadero
.e = error

Las estimaciones desde este enfoque son dependientes de la muestra, por lo que las mediciones hechas para representar el atributo latente tienden a ser inestables (López, 1995). El modelo permite hacer varias presupuestos, pero todos ellos dan forma al aspecto de la confiabilidad de los puntajes. Por lo tanto, las estrategias de análisis psicométrico para obtener las evidencias de confiabilidad, validez y distribucionales fueron se presentan en la Figura 6, que muestran las facetas de validez exploradas para el instrumento bajo estudio, así como los instrumentos usados en cada uno de ellos, los datos de la muestra total en cada año, y el tamaño muestral efectivo con que se obtuvo finalmente los datos usando el instrumento correspondiente.

A. Estrategias para el análisis de la validez

Se aplicaron varias estrategias para evidenciar la validez de las inferencias de los puntajes de la prueba que es objeto de esta investigación. Estas evidencias se refieren a la validez de contenido, de constructo, de la estructura interna y de criterio. Un sumario

	Instrumento	Datos (N total)	N efectivo
Validez de Criterio			
Dificultad y discriminación		2004 (n = 254)	254
		2005 (n = 107)	107
		2006 (n = 187)	187
		2007 (n = 167)	167
Validez de Criterio			
Criterio Concurrente			
Maduracion	Prueba Gestáltica de Brenner	2006 (n = 187)	26
		2007 (n = 167)	80
Validez de Constructo			
Constructo convergente			
Habilidades específicas	Bender – QSS	2004 (n = 254)	
		2005 (n = 107)	107
		2006 (n = 187)	56
	VMI	2007 (n = 167)	87
Constructo Divergente			
Prueba de desarrollo	TEPSI	2007b (n = 30)	30
Inteligencia/Aptitudes	PDP	2007b (n = 30)	30
	DAP:IQ	2006 (n = 187)	180
		2007 (n = 167)	149
	TONI-2	2007 (n = 167)	87
Estructura Interna		2005 (n = 107)	461
		2006 (n = 187)	
		2007 (n = 167)	

CP: Cuestionario para Profesores (Kenny y Chekaluk, 1993)
CINR: Cuestionario de identificación de niños en riesgo: (Cadieux y Boudreault, 2002)
PRA: Programa de Recuperación Académica.

Figura 6
Relación entre los instrumentos, tamaño muestral y facetas de validez exploradas

A1. Validez de contenido (Objetivo 1 de investigación)

Análisis de ítems. En el análisis de la validez del contenido, Hay métodos racionales y cuantitativos para evaluar el contenido (AERA, APA, & NCME, 1999); cuantitativamente, el contenido de los ítems fue evaluado por un análisis de ítems, que tradicionalmente incluye la dificultad, discriminación y homogeneidad de los ítems.

a. ***Discriminación.*** La identificación de la discriminación de un ítem permite identificar si el mismo puede diferenciar niños con rendimiento bajos y alto. Para la discriminación de los ítems, se usó la correlación ítem test corregida por espuriedad (Thorndike, 1980). Anastasi y Urbina (1997) y Garret (1971) señalan que índices de discriminación con una magnitud aún de 0.20 son cuantitativamente aceptables en algunas circunstancias, pero las recomendaciones más populares indican que una magnitud mínima de 0.30 es un nivel aceptable (Thorndike, 1980; Nunnally & Bernstein, 1995). Ya que los ítems individuales tienden a tener bajas correlaciones con otras variables contínuas (Nunally & Bernstein, 1995), para tomar decisiones sobre la validez de los ítems la presente investigación usó un límite inferior es 0.25, de tal modo que se pone un compromiso la elección de buenos ítems discriminadores y la falta de moderada confiabilidad de los ítems.

b. ***Dificultad.*** La dificultad del ítem se determinó por el porcentaje de examinados quienes responden correctamente al ítem. La evaluación de los niveles apropiados de dificultad usó el criterio de Anastasi y Urbina (1997), en que se considera óptima una dificultad promedio de 0.50, e ideal una distribución de ítems entre 0.15 y 0.85, pero ya que las pruebas de despistaje se indican para detectar a niños con futuros problemas, entonces los ítems deberían tener un nivel de dificultad cercano a la tasa de selección deseada

(Anastasi & Urbina, 1997) ya que una prueba de despistaje generalmente tiene el propósito de ayudar a identificar el 10% más bajo de la población (Mathews, 1986), los ítems deberían ser respondidos correctamente por el 90% de la muestra. Los ítems más frecuentes que el 90% añaden poco valor discriminativo, tanto así como los ítems incorrectamente respondidos por menos del 40% de la población (A. Kline, comunicación personal, Marzo 1980, citado en Simmons, 1988). Fuera del rango de dificultad entre 0.40 y 0.90, los ítems podrían ser cuestionables para propósitos de despistaje (Simmons, 1988); por lo tanto, la presente investigación usó como criterio este rango de dificultad.

c. ***Homogeneidad de los ítems*.** La homogeneidad de los ítems se refiere al grado en que las correlaciones entre ellas son adecuadas para definir un grupo que homogéneamente mide el constructo pretendido, es decir, que correlacionan similarmente (Clark & Watson, 1995). La homogeneidad de expresa mediante la correlación inter-ítem promedio dentro del área o constructo de interés (Cronbach, 1951. Para evaluar la homogeneidad se han propuesto varias recomendaciones (Clark & Watson, 1995; Briggs & Check, 1986), entre las que destacan que un rango general entre 0.15 y 0.50 para constructos de espectro reducido, y un rango entre 0.15 y 0.20 para constructos de amplio espectro y de orden superior.

La estrategia de análisis fue como se ilustra en la Figura 7. Los pasos secuenciales examinarán, respectivamente, a) la variabilidad de los parámetros de los ítems en la muestra de los colegios en cada año (intra-año), b) la variabilidad entre los años (entre-años), y c) la adecuación según los estándares. Cada fase dio soporte a distintos aspectos

del funcionamiento de los ítems, siendo la comprobación fundamental la invariabilidad de estos indicadores, es decir, la asunción que los parámetros de dificultad, discriminación y homogeneidad son invariables a través de las diferentes muestras de niños y a través del tiempo; y que las potenciales variaciones fueron por eventos aleatorios o únicos en cada grupo y momento examinado. Esta confirmación supondrá que los ítems funcionan similarmente en estas diferentes condiciones, y que la estimación en la muestra total de estos parámetros puede representar la homogeneidad y estabilidad parámetro, y consecuentemente, tomar decisiones sobre lo adecuabilidad dentro del instrumento. En el análisis intra-año, en que se comparó la dificultad del ítem entre los colegios (p_1 y p_2) dentro de cada año, se aplicó la transformación arcoseno ($2\left(arcoseno\left(\sqrt{p}\right)\right)$) a p_1 y p_2 (Cohen, 1988), con el fin de normalizar la distribución binomial correspondiente de p_1 y p_2 y mejorar el poder estadístico de las diferencias (Cohen, 1988); en la comparación entre-años, se aplicó una prueba χ^2 de bondad de ajuste contra la distribución equiprobable, con 3 grados de libertad; finalmente se estableció una corrección Bonferroni al nivel alfa de 0.05 (0.05/*k*, en que *k* es el número de pruebas estadísticas). La magnitud del efecto de las diferencias se dará mediante la *V de Cramer*, teniendo en cuenta los niveles cualitativos sugeridos por Cohen (1988): 0.10, 0.30 y 0.50 para un efecto pequeño, moderado y grande, respectivamente. Para la comparación de los índices de discriminación (correlaciones ítem-test corregida), en el nivel intra-año, se aplicó una prueba z de diferencias de correlaciones independientes; para el nivel entre-años, de aplicó una prueba (Q) para comparar ≥3 correlaciones independientes. Como en el análisis de las diferencias, se ajustó el valor *p* de acuerdo al enfoque Bonferroni.

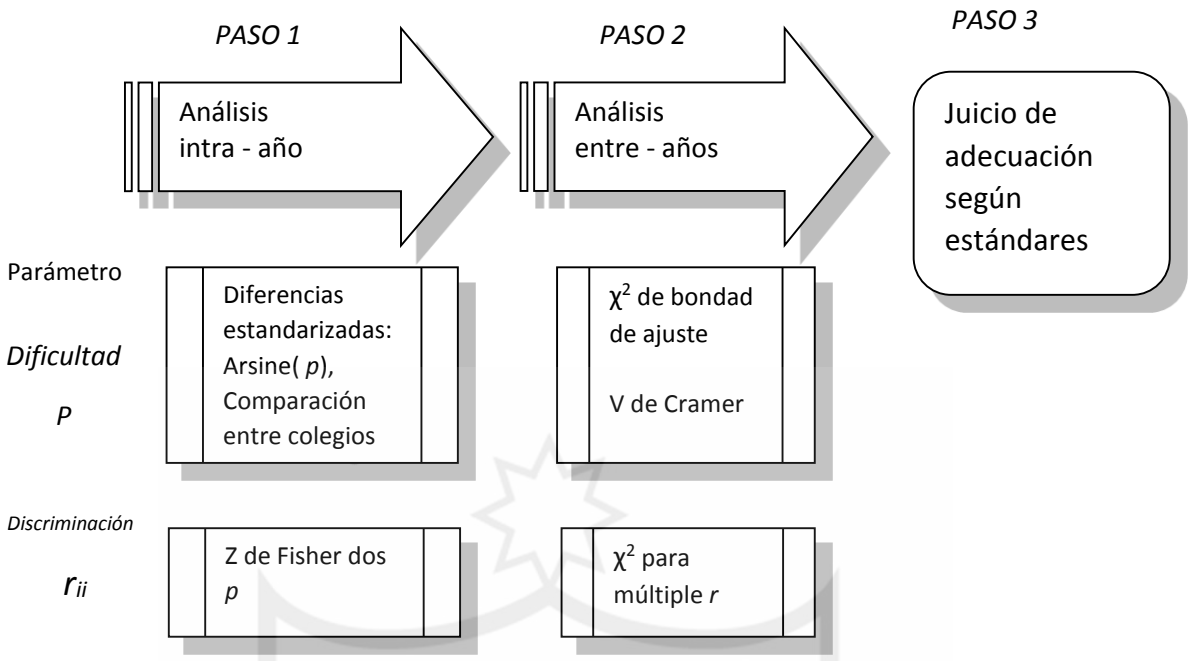


Figura 7
Secuencia del análisis de ítems para evidenciar cuantitativas de validez de contenido

A2. Validez de la Estructura Interna (Objetivo 2 de investigación)

Para examinar la dimensionalidad, se aplicó un enfoque de ecuaciones estructurales (SEM, Structural Equation Modeling) mediante un análisis factorial confirmatorio (CFA, Confirmatory Factor Analysis). Con esta metodología se probarán varios modelos de medición para hallar el que mejor represente la estructura del constructo. La unidad de análisis fueron grupos (clusters) de ítems, cuyos criterios de agrupamiento y distribución final se describirán en las secciones correspondientes al análisis de la dimensionalidad. Este agrupamiento ayuda a evitar problemas de convergencia en el CFA y hacer parsimonioso en análisis cuando se tienen muchos ítems (Byrne, 2006). En esta tabla presenta subescalas compuestas por grupos de 2 o 6 ítems, agrupadas racionalmente por su contenido. Preliminarmente, se evaluaron la normalidad univariada y multivariada

mediante la estandarización de los índices de simetría y curtosis de Fisher (Hopkins & Weeks, 1990); ambos se obtuvieron por el programa EQS (Bentler & Wu, 2004).

Método de ajuste

Se aplicó el método de máxima verosimilitud (Maximum Likehood), un método usual para modelar los datos. Sin embargo, las unidades primarias de análisis son variables dicotómicas, que generalmente muestran grandes asimetrías y curtosis, relaciones no lineales y producen factores espurios (Bernstein & Teng, 1989; Curran, West & Finch, 1995; De Rbuin, 2004). Este problema afecta comúnmente la correcta estimación del χ^2, tendiendo a producir infravaloración del error estándar de los parámetros de interés (Curran et al., 1995; De Rbuin, 2004). Para atenuar estos problemas, primero se agruparán los ítems en parcelas antes de someterlos al CFA. En contraste con otros métodos, las parcelas agruparán ítems de acuerdo al contenido (De Bruin, 2004), lo que indica que de manera a priori de formarán escalas homogéneas referidas a las habilidades específicas que pretenden medir los ítems agrupados. Esto es más coherente teóricamente dado que los ítems construidos y seleccionados en BDPG representan contenidos específicos, y que aportan varianza significativa dentro de cada cluster. Sin embargo, se presupone que hay un factor dominante que explicaría la varianza la varianza común entre las parcelas.

Para atenuar el efecto de la no normalidad de los datos sobre el método de ajuste, se aplicó la corrección Satorra-Bentler (SB-χ^2; Satorra & Bentler, 1994; 2001), que re-escala el estadístico χ^2. Todos los cálculos del CFA se efectuarán mediante el programa EQS (Bentler & Wu, 2004).

Índices de ajuste del modelo

Para la evaluación del ajuste se usaron índices absolutos y relativos; los primeros basados en la discrepancia entre las covarianzas predichas y las ajustadas (RMSEA <= 0.05, SRMR <=0.05), y los segundos de ajuste comparativo con el modelo nulo (CFI >= 0.95, NNFIT >= 0.95); estos últimos parecen mejor recomendados en análisis con muestras pequeñas debido a su falta de sensibilidad al tamaño muestral (Bentler, 1990; Fan, Thompson & Wang, 1999; Marsh, Balla & McDonald, 1988). Los índices de ajustes usaron los resultados reescalados obtenidos con el método de Satorra-Bentler.

A3. Validez convergente/divergente (Objetivo 3 de investigación)

i. *Relaciones convergentes.*

Procedimiento. Este aspecto de la validez cubre las relaciones teóricas entre constructos, en las que se espera que las correlaciones sean de mayor o menor magnitud, de acuerdo a lo que teóricamente de suponen que deben compartir un monto de varianza entre los constructos.

Instrumentos. Se seleccionaron pruebas visomotoras y de desarrollo aptitudinal escolar general. Sobre las primeras, la versión modificada de la Prueba Gestáltica de Bender evaluada por el Sistema Brannigan y Brunner (2002), y la Prueba de Integración Visomotora, VMI (Beery, 2000). Sobre la prueba de desarrollo de aptitudes para el aprendizaje escolar, el instrumento fue la Prueba de Diagnóstico Preescolar (PDP; De la Cruz, 1988)

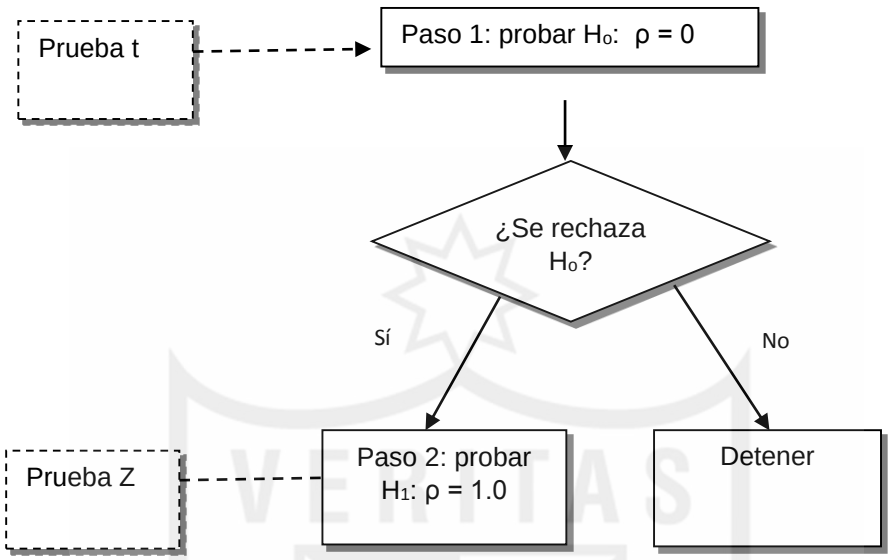
Análisis estadístico. Se usaron correlaciones *r* de Pearson, asumiendo que se cumplan los presupuestos estadísticos para su aplicación. Se aplicaran dos enfoques respecto a la validez convergente: uno probó la tradicional hipótesis nula (H_o: $r = 0.0$), y el otro enfoque probó la máxima correlación posible tomando en cuenta las confiabilidades de cada medida (H_i: $r = 0.0$). Este análisis fue conducido secuencialmente, ya que primero se probará que las relaciones no se obtienen por azar, y seguidamente, se probará si las relaciones alcanzan su máxima magnitud considerando las confiabilidades conjuntas entre el PDP y el BDPG (Braden 1986). En la Figura 8, se observa la secuencia de dos pasos para efectuar el análisis.

ii. *Relaciones divergentes.*

Procedimiento. Se evaluó las relaciones de los constructos medidos por la batería con otros constructos con los cuales se espera que la magnitud sea baja o cero. Estas relaciones divergentes indicarían que la batería mide aspectos no cubiertos que son consistentes con la finalidad y el diseño de construcción del BDPG.

Instrumentos. Estas pruebas de desarrollo general para niños (Test de Desarrollo Psicomotor, TEPSI; Haeussler & Marchant, 2003), y de inteligencia no-verbal (Dibujo de la Figura Humana, DAP: IQ; Reynolds & Hickman, 2004).

Análisis estadístico. Se usaron correlaciones r de Pearson, asumiendo que se cumplan los presupuestos estadísticos para su aplicación.



Nota: El valor de la hipótesis alternativa en el paso dos, ρ = 1.0, se obtiene de la máxima confiabilidad conjunta $r_{xxyy} = \sqrt{r_{xx} r_{yy}}$

Figura 8
Pasos para aplicar el enfoque de correlaciones en pequeñas muestras (Braden, 1986) para el análisis de las relaciones convergentes.

A4. Validez por Diferenciación de grupos extremos (Objetivo 4 de investigación)

Procedimiento. Se hizo mediante la identificación de *Grupos conocidos* (Thorndike, 1989). Se efectuará la comparación entre grupos identificados en los extremos de la distribución del puntaje en el criterio. El criterio fue la percepción de la profesora respecto al rendimiento académico de los estudiantes. Se aplicó el siguiente procedimiento:

Se solicitará a la profesora que elija a 5 alumnos del más alto y del más bajo rendimiento académico. Este número corresponde al 23% de la muestra total de niños de primer grado en el colegio, y es un compromiso entre la

sugerencia para maximizar el poder en este diseño de grupos extremos (D'Agostino, & Cureton, 1975) y la elección de alrededor de 1 desviación estándar usada en otras investigaciones (Cadieux & Boudreault, 2002; Flynn & Rahber, 1994). Su usaron las siguientes instrucciones:

> *Por favor elija a los 5 alumnos de mejor rendimiento académico, teniendo en cuenta el rendimiento en general pero sin considerar el comportamiento en el aula. También, elija a los 5 alumnos de más bajo rendimiento académico en general, pero sin tomar en cuenta la conducta en el aula. Si estos alumnos tienen mal comportamiento pero rinden aceptablemente o mejor, entonces no los elija.*

Se administrarán las pruebas en orden balanceado, para controlar los posibles efectos del orden de presentación de las tareas de las pruebas (Thorndike, 1989).

Instrumentos. Las áreas evaluadas son de a) integración visomotora (Prueba Gestáltica de Bender Modificado (Brannigan & Brunner (2002); de disposición perceptual general para el aprendizaje (Prueba Gestáltica Antón Brenner, Woodburn & Boschini, 1993); de percepción visual (Test de Percepción Visual no Motriz, TPVNM; Colarusso & Hammill, 1980); y de inteligencia no verbal (Dibujo de la Figura Humana para Estimación Intelectual en Niños, Adolescentes y Adultos: DAP: IQ; Reynolds & Hickman, 2004).

Análisis estadístico. Se utilizaron dos métodos: la correlación *r* modificado por Feldt (1961) y por Peters (1941; Thorndike, 1989). Las diferencias fueron evaluadas en términos de *d de Cohen*, derivado de las estimaciones *r* anteriores. Para atenuar el impacto de trabajar con grupos extremos (y

restricción del rango consecuente), el valor r obtenido fue ajustado por el efecto de segmentación o restricción del rango (Preacher, 2007; Preacher, Rucker, McCallum & Nicewander, 2005) (ver el ANEXO C). Finalmente, se derivara el coeficiente de magnitud del efecto d (Cohen, 1988), corrigiéndolo por el potencial sesgo positivo originado por el pequeño tamaño muestral (Hedges & Olkin, 1985):

$$g = d\left(1 - \frac{3}{4N - 9}\right)$$

A5. Validez de criterio concurrente (Objetivo 5 de investigación)

Procedimiento. Otra estrategia de validez con criterios de hizo mediante el grado de relación de los puntajes de BDGPG con otros instrumentos que pretenden medir similares o el mismo constructo (Thorndike, 1989). La estrategia consiste en aplicar los instrumentos que se correlacionaran en el mismo periodo de tiempo. Se aplicaron en las mismas condiciones ambientales y de administración para todos los niños. La muestra en este procedimiento se obtuvo en el 2006 (n = 26) y 2007 (n = 80).

Instrumentos. El instrumento para la validación concurrente fue *Prueba Gestáltica Antón Brenner* (PGAB; Woodburn & Boschini, 1993), que evalúa el aprestamiento desde un enfoque holístico, enfatizando la habilidad perceptual visual.

Análisis estadístico. Se usó la correlación Pearson para estimar la asociación lineal entre la BDPG y PGAB, independientemente en el grupo del 2006 y 2007. Para obtener una estimación menos sesgada de la correlacion en la muestra del 2006 (n = 26), que es la que tiene un tamaño muestral pequeño, se aplicó el ajuste de Olkin y Pratt (1958):

$$\rho = r\left[1 + \frac{\left(1 - r^2\right)}{2\left(n - 3\right)}\right]$$

Este estimador permite obtener un mejor control del sesgo comparado con otros procedimientos (Zimmerman, Zumbo & Williams, 2003).

B. Estrategias para el análisis de la confiabilidad (Objetivo 6 de investigación)

Consistencia interna.

Procedimiento. Se aplicaron en una sola ocasión la prueba BDPG al grupo de niños examinados, siguiendo las instrucciones estandarizadas. Los datos obtenidos en esta aplicación sirvieron para todos los análisis esenciales descritos en esta investigación.

Análisis. Se aplicaron el coeficiente KR-20, dado que este coeficiente se usa para ítems dicotómicos (Thorndike, 1989). Ya que la confiabilidad por este método puede restringir su amplitud debido a las diferencias en las dificultades de los ítems (Horst, 1953; Charter, 1995), se aplicará la corrección de Horst (Horst, 1953), asumiendo que el modelo de medición de los ítems que subyace al KR-20 es satisfecho (Merino & Charter, en prensa). Para saber lo adecuado de la confiabilidad por consistencia interna obtenida, no hay criterios uniformes y varían de acuerdo a varios autores (Charter, 2003). Teniendo en cuenta los criterios de Cicchetti (Cicchetti & Sparrow, 1981; Cicchetti, 1994), Nunally y Bernstein (1995), y Anastasi y Urbina (1997), así como la revisión empírica de Charter (2003), se estableció un criterio mínimo de 0.75. Para las comparaciones entre los coeficientes de consistencia interna, se usó el enfoque de prueba de hipótesis basados en los trabajos de Feldt, para la comparación entre muestras independientes (Feldt, Woodruff & Salih, 1987), y viabilizado por el programa ALPHATST (Lautenschlager, 1989; Merino & Lautenschlager, 2003)

Las estimaciones de confiabilidad fueron analizadas en dos niveles: intra-año y entre-años. En la Figura 9 se muestra esta secuencia. En el nivel intra-año, se compararán las confiabilidades para evaluar las diferencias entre las muestras. Luego, se pasará a evaluar las diferencias entre los años. Ambos análisis probarán la constancia de las confiabilidades en estas dos condiciones.

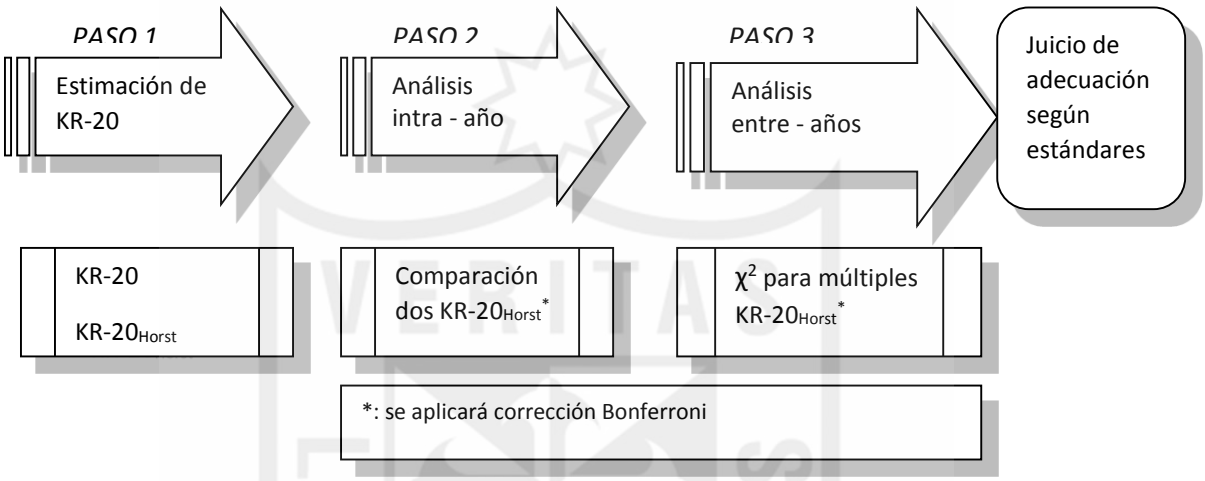


Figura 9
Secuencia de análisis para las estimaciones de consistencia interna

C. Estrategias de análisis para identificar el piso y techo del puntaje, y la gradiente del ítem en la distribución del puntaje total y de las subescalas (Objetivo 7)

Se siguieron las recomendaciones descritas en Bracken (2000) y Rathvon (2004). Estos criterios establecen la relación entre el máximo (techo) y mínimo (piso) puntaje posible y el puntaje directo en ±2 desviaciones estándares de la media. Esto permite evaluar el grado en que la distribución de puntajes se extiende más allá de este punto, en que se puede identificar rendimientos extremos que requieran no evaluación o intervención en la mayor brevedad posible. Por otro lado, la gradiente del ítem relaciona el cambio entre cada puntaje directo y sus correspondientes cambios entre cada puntaje

estandarizado. La aplicación de estos criterios es cualitativa, en que mediante la observación se determina si se cumple el criterio del techo, piso o gradiente del ítem.

D. Estrategias para el desarrollo de la información normativa y la interpretación estandarizada y normativa de los puntajes del BDPG (Objetivo 8 de investigación)

El examen de las diferencias entre las características de la muestra, respecto al género, periodo y el colegio de procedencia de los niños, servió para evaluar la construcción de distribuciones unificadas de cada periodo de evaluación, y así obtener puntajes estandarizados. Una vez establecido que las diferencias no son estadísticamente significativas o que estas presentan baja magnitud, las distribuciones unificadas fueron ajustadas a una distribución teórica que sea menos dependiente de variaciones de muestreo y se acerque a la distribución de puntajes verdaderos.

Las distribuciones de puntajes de las subescalas y del puntaje total se ajustarán al *modelo beta binomial compuesto de 4 parámetros* (4PB; Hanson, 1991; Lord, 1965), considerando que este modelo representa mejor la forma distribucional de los puntajes provenientes de pruebas de rendimiento estandarizado, como se mencionó en la introducción. Luego del ajuste, se comprobará si el ajuste ha sido satisfactorio mediante la prueba χ^2 de bondad de ajuste. Con la aplicación de este modelo, se obtendrá la distribución de puntajes verdaderos que fue la base para la obtención de normas preliminares. El procedimiento de ajuste se hizo mediante el programa BB-CLASS (Brennan, 2004), que requiere la información paramétrica de los momentos de la los datos directos de los puntajes; estos indicadores son cuatro: dos indicadores de la forma de la distribución, α y β, y el límite superior e inferior estimados de la distribución del puntaje verdadero. El resultado de este ajuste se informará mediante los parámetros obtenidos, la

proporción y frecuencia ajustada según la distribución teórica aplicada, la prueba χ^2 de bondad de ajuste comparando la distribución original y ajustada (Brennan, 2004), y finalmente *gráficos kernel* en el que se superpondrán la densidad empírica y ajustada. Para cuantificar el grado de ajuste, se aplicó la ratio χ^2 / grados de libertad (g.l.), en que los valores cercanos a 1 y debajo de 2 son considerados un buen ajuste (Gempp & Cuesta, 2007).

E. Estrategias para identificar el impacto de las diferencias demográficas (sexo del niño, colegio de procedencia) sobre el rendimiento en la Batería de Despistaje para Primer Grado (Objetivo 9 de investigación)

Las diferencias fueron evaluadas mediante un análisis multivariado de la varianza (MANOVA), teniendo como variables dependientes los puntajes de las subescalas, y como variables de efecto el género y el colegio de procedencia. Se eligió este modelo multivariado porque las variables dependientes (los puntajes de las subescalas) claramente forman un sistema relacionado de variables entre las que subyace un constructo de habilidad general (Rencher, 2002) y que es estimable por el puntaje total. El modelo de multivariado aplicado es completamente factorial, que incluye los efectos principales fijos (género y colegio) y la interacción entre estas variables (género x colegio). Este modelo se representa de la siguiente manera:

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \varepsilon_{ijk}$$

En que y_{ijk} son los puntajes en las variables dependientes, μ es la media general, α_i es el es el efecto principal de una de las variables demográficas (sexo), β_j es

el efecto principal de la otra variable demográfica (colegio), γ_{ij} el efecto de la interacción de ambas, y ε_{ijk} el error o residual.

Ya que el tamaño de la muestra entre los grupos es aproximadamente similar, los efectos de las diferencias en la varianza sobre el error Tipo II fue menor (Rencher, 2002). Primero, se aplicó una prueba *F* para examinar en si el modelo multivariado explica los efectos de las variables demográficas sobre cada variable dependiente. Si se detectan efectos sistemáticos debajo del nivel nominal de significancia (0.05), el siguiente paso fue examinar el origen de los efectos estadísticamente significativos en las variables demográficas. El estimador multivariado para este análisis fue λ (*Lambda de Wilks*), desde el que se transformará a T^2 *de Hotelling* mediante la modificación de Rencher (2002). La T^2 *de Hotelling* es una apropiada extensión natural de la *t de Student* en la comparación de dos muestras sobre un sistema de variables Rencher (2002). El estimador de magnitud del efecto fue η^2 (eta cuadrado). Las recomendaciones sobre la descripción cualitativa de este índice sugieren los siguientes parámetros de interpretación: pequeña = 0.01; moderada = 0.06; grande = 0.14 (Cohen, 1988; Sink y Stroh, 2006).

CAPÍTULO 3
RESULTADOS

VALIDEZ DE CONTENIDO (ANÁLISIS DE ÍTEMS)

Dificultad del ítem

Dificultad del ítem en Conocimiento de letras y palabras. La magnitud promedio de las diferencias estandarizadas de la dificultad del ítem fue 0.09 y -0.11 en los años 2006 y 2007, respectivamente (Tabla 7). Por lo tanto, estas diferencias entre la dificultad de los ítems fueron pequeñas e insignificantes en estos años, indicando que la variación en la dificultad del ítem no tendrían que ser interpretables. La excepción ocurrió en el año 2004, ya que el 77.2% de los ítems mostraron grados de dificultad mayor o igual a 0.30 (d promedio = 0.35) y sugieren una diferencia moderada; y estas diferencias favorecieron al colegio MI. Dado que las diferencias intra-año fueron principalmente pequeñas, y aquellas moderadas entre unos años parece un evento específico y no repetible, se tomaron las muestran conjuntas en cada año para probar la hipótesis de igualdad entre los años (análisis entre-años).

En la

Tabla 8 se muestran los resultados de la comparación entre años para la subescala Letras y Palabras. Se hizo aplicó un χ^2 y una estimación de magnitud del efecto correlacional (V_{cramer}) a cada ítem a través de los años. Se halló que 4 de los ítems mostraron frecuencias de respuesta estadísticamente diferentes entre los años. En estos ítems, hubo más respuestas correctas en el año 2007, pero la magnitud de estas diferencias no superó el valor V_{cramer} igual 0.20 y la frecuencia aún se mantuvo entre los niveles esperados. En los demás ítems, las diferencias entre las dificultades de los ítems fueron por el error de muestreo, y la magnitud se mantuvo cercana a cero. Respecto a la escala Conocimiento de Letras y Palabras, podemos afirmar que la dificultad de los ítems fue poco variable y se puede aceptar un valor común representado por la dificultad promedio ($\bar{X}_p$).

Tabla 7
Subescala Letras y Palabras: Diferencia arcsine en la dificultad de los ítems, entre los colegios en cada año muestrado

	2004			2007			2006		
	MI	SM	d[a]	MI	SM	d[a]	MI	SM	d[a]
P8	0.96	0.77	0.60	0.95	0.95	0.0	0.86	0.84	0.06
P9	0.79	0.56	0.50	0.72	0.66	0.13	0.62	0.69	-0.15
P10	0.53	0.39	0.28	0.56	0.53	0.06	0.30	0.44	-0.29
P11	0.59	0.49	0.20	0.79	0.69	0.23	0.52	0.65	-0.26
P12	0.49	0.34	0.31	0.40	0.31	0.19	0.32	0.34	-0.04
P13	0.59	0.33	0.53	0.65	0.51	0.28	0.32	0.52	-0.41
P14	0.61	0.44	0.34	0.71	0.66	0.11	0.58	0.60	-0.04
P15	0.77	0.59	0.39	0.78	0.72	0.14	0.65	0.75	-0.22
P16	0.50	0.33	0.35	0.36	0.26	0.22	0.33	0.31	0.04
P17	0.55	0.33	0.45	0.42	0.38	0.08	0.43	0.44	-0.02
P18	0.44	0.29	0.31	0.48	0.38	0.20	0.30	0.28	0.04
P19	0.47	0.31	0.33	0.50	0.46	0.08	0.35	0.41	-0.12
P48	0.38	0.29	0.19	0.38	0.33	0.10	0.34	0.28	0.13
P49	0.47	0.29	0.37	0.52	0.56	-0.08	0.32	0.35	-0.06
P50	0.52	0.43	0.18	0.55	0.61	-0.12	0.48	0.56	-0.16
P51	0.52	0.41	0.22	0.61	0.56	0.10	0.43	0.53	-0.20
P52	0.60	0.39	0.42	0.59	0.63	-0.08	0.48	0.55	-0.14
P53	0.42	0.22	0.43	0.38	0.31	0.15	0.29	0.32	-0.07
Máx.	0.77	0.59	0.45	0.78	0.72	0.22	0.65	0.75	0.13
M.	0.52	0.36	0.33	0.52	0.49	0.12	0.42	0.45	0.10
Mín.	0.38	0.22	0.18	0.36	0.26	-0.12	0.29	0.28	-0.22

[a]: diferencias estandarizadas basadas en la transformación arcsine de la dificultad de los ítems.
[b]: promedio obtenido de los valores absolutos

Tabla 8
Subescala Letras y Palabras: Dificultad de los ítems para las muestras totales en cada año

	Estimaciones de dificultad					Prueba y magnitud del efecto	
	2004 (n = 189)	2005 (n = 107)	2006 (n = 187)	2007 (n= 167)	$\bar{X}_p$	χ^2	V_{cramer}
p8	0.87	0.84	0.85	0.95	0.88	11.58**	0.13
p9	0.68	0.68	0.65	0.69	0.68	0.65	0.03
p10	0.46	0.46	0.37	0.54	0.46	10.34	0.12
p11	0.54	0.49	0.58	0.74	0.59	22.97**	0.18
p12	0.42	0.41	0.33	0.35	0.38	4.14	0.07
p13	0.47	0.40	0.42	0.57	0.47	10.24	0.12
p14	0.53	0.53	0.59	0.68	0.58	10.20	0.12
p15	0.68	0.67	0.70	0.75	0.70	2.50	0.06
p16	0.42	0.41	0.32	0.31	0.37	6.91	0.10
p17	0.44	0.46	0.44	0.40	0.44	1.01	0.03
p18	0.37	0.23	0.29	0.43	0.33	14.53**	0.14
p19	0.39	0.29	0.38	0.48	0.38	10.18	0.12
p48	0.33	0.34	0.31	0.35	0.33	0.58	0.02
p49	0.38	0.39	0.34	0.54	0.41	15.77**	0.15
p50	0.48	0.50	0.52	0.58	0.52	3.69	0.07
p51	0.47	0.45	0.48	0.59	0.50	7.78	0.10
p52	0.50	0.42	0.51	0.61	0.51	10.08	0.12
p53	0.32	0.24	0.31	0.34	0.30	3.06	0.06
Máx.	0.87	0.84	0.85	0.95	0.88		0.18
M.	0.49	0.46	0.47	0.55	0.49		0.097
Mín.	0.32	0.23	0.29	0.31	0.30		0.02

.a: El nivel α ajustado por Bonferroni (α/n), en que *n* es el número elementos para comparar, es 0.05 / 4 = 0.012; ** p < 0.012

Discriminación Fonológica. Se analizaron las diferencias en la dificultad de los ítems entre los colegios dentro de cada año (Tabla 9). El análisis intra-año demostró nuevamente que las diferencias estandarizadas entre los colegios estuvieron alrededor del nivel considerado como bajo y pocos cerca de tener diferencias moderadas. Las diferencias estandarizadas promedio entre los colegios para los años 2004, 2006 y 2007 fueron, respectivamente, 0.22, 0.013 y 0.07.Adicionalmente, para aquellos ítems que mostraron diferencias moderadas, sus valores estuvieron dentro del mismo rango de aceptabilidad. Entonces, estas diferencias fueron atenuadas por el nivel esperado de dificultad apropiada para la presente investigación. La relativa y aceptable constancia

intra-año de la dificultad de los ítems nos lleva a analizar las diferencias entre-años sin separar la muestra por colegios; estos resultados se muestran en la Tabla 10.

Al comparar la dificultad del os ítems entre los años (Tabla 10), hallamos que la dificultad de 6 ítems entre los años fue variable en términos de significancia estadística, pero ninguno supero un V_{cramer} igual a 0.19. Ya que las diferencias en en los otros ítems no fueron estadísticamente significativos, esta variación se explicó por el error de muestreo, y la magnitud de estas diferencias en términos de V_{cramer} fue cercano a cero. El valor promedio de V_{cramer} fue 0.09. Como en el análisis de la subescala anterior, el promedio ($\bar{X}_p$)de la dificultad de todos los años puede representar exitosamente este parámetro psicométrico en cada ítem.

Tabla 9
Subescala Disc. Fonológicas: Diferencia arcsine en la dificultad de los ítems, entre los colegios en cada año muestreado

	2004			2006			2007		
	MI	SM	d[a]	MI	SM	d[a]	MI	SM	d[a]
p20	0.73	0.64	0.19	0.53	0.57	-0.08	0.72	0.70	0.04
p21	0.76	0.60	0.35	0.68	0.59	0.19	0.72	0.71	0.02
p22	0.72	0.57	0.32	0.57	0.47	0.20	0.68	0.51	0.35
p23	0.78	0.67	0.25	0.63	0.81	-0.41	0.79	0.79	0.00
p24	0.47	0.33	0.29	0.27	0.30	-0.07	0.50	0.44	0.12
p25	0.63	0.60	0.06	0.77	0.70	0.16	0.81	0.76	0.12
p26	0.36	0.22	0.31	0.37	0.44	-0.14	0.35	0.43	-0.16
p27	0.43	0.28	0.32	0.37	0.33	0.08	0.35	0.23	0.27
p28	0.52	0.48	0.08	0.59	0.54	0.10	0.68	0.61	0.15
p29	0.43	0.32	0.23	0.36	0.33	0.06	0.42	0.30	0.25
p30	0.55	0.41	0.28	0.48	0.49	-0.02	0.59	0.51	0.16
p31	0.47	0.33	0.29	0.29	0.22	0.16	0.36	0.32	0.08
p32	0.80	0.65	0.34	0.65	0.75	-0.22	0.76	0.84	-0.20
p33	0.75	0.70	0.11	0.62	0.63	-0.02	0.69	0.72	-0.07
p34	0.74	0.83	-0.22	0.71	0.71	0.00	0.79	0.74	0.12
p35	0.63	0.45	0.36	0.34	0.28	0.13	0.40	0.41	-0.02
p36	0.52	0.35	0.34	0.29	0.24	0.11	0.39	0.36	0.06
Máx.	0.8	0.83	0.36	0.77	0.81	0.2	0.81	0.84	0.35
M.	0.61	0.50	0.23	0.50	0.49	0.01	0.59	0.55	0.08
Mín.	0.36	0.22	-0.22	0.27	0.22	-0.41	0.35	0.23	-0.2

[a]: diferencias estandarizadas basadas en la transformación arcsine de la dificultad de los ítems.
[b]: promedio obtenido de los valores absolutos

Tabla 10
Subescala Disc. Fonológicas: Dificultad de los ítems para las muestras totales en cada año

	Estimaciones de dificultad					Prueba[a] y magnitud del efecto	
	2004	2005	2006	2007	$\bar{X}_p$	χ^2	V_{cramer}
p20	0.69	0.68	0.55	0.71	0.66	12.66**	0.13
p21	0.68	0.70	0.64	0.72	0.69	2.60	0.06
p22	0.64	0.64	0.52	0.59	0.60	6.81	0.10
p23	0.73	0.74	0.72	0.79	0.75	2.57	0.06
p24	0.40	0.42	0.28	0.47	0.39	14.56**	0.14
p25	0.61	0.61	0.73	0.78	0.68	16.89**	0.16
p26	0.29	0.34	0.41	0.39	0.36	6.96	0.10
p27	0.36	0.30	0.35	0.29	0.33	2.84	0.06
p28	0.50	0.40	0.56	0.64	0.53	16.43**	0.15
p29	0.37	0.36	0.35	0.36	0.36	0.22	0.01
p30	0.48	0.54	0.49	0.54	0.51	1.85	0.05
p31	0.40	0.36	0.25	0.34	0.34	10.10	0.12
p32	0.73	0.74	0.70	0.80	0.74	5.00	0.08
p33	0.72	0.69	0.63	0.71	0.69	4.18	0.08
p34	0.78	0.79	0.71	0.76	0.76	3.39	0.07
p35	0.54	0.44	0.31	0.41	0.43	20.58**	0.17
p36	0.44	0.34	0.26	0.37	0.35	13.27**	0.14
Máx.	0.78	0.79	0.73	0.8	0.76		0.17
M.	0.55	0.53	0.50	0.57	0.54		0.10
Mín.	0.29	0.3	0.25	0.29	0.33		0.01

[a]: El nivel α ajustado por Bonferroni (α/n), en que n es el número elementos para comparar es 0.05 / 4 = 0.012; **: p < 0.01

Percepción Visual. En el análisis intra-año, las diferencias entre los colegios fueron pequeñas (

Tabla 11), pues la diferencia media de la dificultad fue por lo general < 0.18. Los valores promedio de las diferencias entre los años 2004, 2006 y 2007 fueron 0.13, 0.09 y 0.14, respectivamente. La magnitud de estas diferencias estandarizadas está centradas en niveles bajos, de tal modo que puede establecerse que las diferencias entre los colegios respecto a los ítems de estas subescala tienden a ser pequeñas. Algunos ítems en particulares años mostraron diferencias moderadas, pero esto fue más bien eventual ya que no se repitió en los demás años. Ya que las diferencias no fueron estadísticamente importantes ni de gran magnitud, se efectuó el análisis entre-años tomando la dificultad de la muestra total en cada año.

La

Tabla 12 exhibe estos resultados. La prueba χ^2 mostró que 7 ítems variaron sistemáticamente su nivel dificultad (ítems 44, 45, 56, 57, 58, 60, 63). Examinando la tabla, en el año 2007 estos ítems tendieron a ser más fáciles, pero su proporción de respuesta no fue tan grande como para darle significancia práctica. La confirmación de ello se obtuvo por los valores V_{cramer},que fueron <0.20. Las dificultades de demás ítems fueron relativamente constantes, y se puede considerar que las fluctuaciones se debieron al error de muestreo.

Tabla 11
Escala: Hab. Percepción Visual: Diferencia arcsine en la dificultad de los ítems, entre los colegios en cada año muestrado

	2004			2006			2007		
	Mi	SM	d[a]	Mi	SM	d[a]	Mi	SM	d[a]
p37	0.6	0.57	0.06	0.57	0.43	0.28	0.64	0.55	0.18
p38	0.61	0.6	0.02	0.53	0.60	-0.14	0.66	0.64	0.04
p39	0.56	0.53	0.06	0.47	0.47	0.00	0.50	0.52	-0.04
p40	0.52	0.54	-0.04	0.50	0.52	-0.04	0.56	0.52	0.08
p41	0.71	0.73	-0.04	0.61	0.67	-0.13	0.72	0.61	0.23
p42	0.71	0.61	0.21	0.61	0.60	0.02	0.65	0.69	-0.09
p43	0.49	0.44	0.10	0.60	0.58	0.04	0.70	0.55	0.31
p44	0.63	0.59	0.08	0.70	0.76	-0.14	0.81	0.75	0.15
p45	0.71	0.53	0.37	0.60	0.59	0.02	0.68	0.72	-0.09
p46	0.55	0.33	0.45	0.54	0.44	0.20	0.50	0.39	0.22
p47	0.73	0.73	0.00	0.79	0.88	-0.24	0.76	0.83	-0.17
p54	0.76	0.82	-0.15	0.79	0.81	-0.05	0.80	0.86	-0.16
p55	0.53	0.68	-0.31	0.60	0.67	-0.15	0.75	0.70	0.11
p56	0.68	0.8	-0.28	0.84	0.75	0.22	0.71	0.63	0.17
p57	0.51	0.69	-0.37	0.74	0.75	-0.02	0.80	0.84	-0.10
p58	0.64	0.73	-0.19	0.84	0.82	0.05	0.81	0.85	-0.11
p59	0.6	0.57	0.06	0.59	0.56	0.06	0.65	0.60	0.10
p60	0.66	0.67	-0.02	0.79	0.77	0.05	0.79	0.89	-0.28
p61	0.54	0.57	-0.06	0.59	0.61	-0.04	0.66	0.59	0.14
p62	0.39	0.41	-0.04	0.43	0.42	0.02	0.42	0.44	-0.04
p63	0.68	0.69	-0.02	0.77	0.73	0.09	0.78	0.87	-0.24
Máx.	0.76	0.82	0.45	0.84	0.88	0.22	0.81	0.89	0.31
M.	0.61	0.62	0.13[b]	0.68	0.68	0.09[b]	0.71	0.70	0.14[b]
Mín.	0.39	0.33	-0.37	0.43	0.42	-0.24	0.42	0.39	-0.28

[a]: diferencias estandarizadas basadas en la transformación arcsine de la dificultad de los ítems.
[b]: promedio obtenido de los valores absolutos

Tabla 12
Escala: Hab. Percepción Visual: Dificultad de los ítems para las muestras totales en cada año

	Estimaciones de dificultad					Prueba[a] y magnitud del efecto	
	2004	2005	2006	2007	$\bar{X}_p$	χ^2	V_{cramer}
p37	0.59	0.49	0.5	0.59	0.54	6.08	0.09
p38	0.61	0.56	0.57	0.65	0.60	3.27	0.07
p39	0.54	0.5	0.47	0.51	0.51	1.80	0.05
p40	0.53	0.59	0.51	0.54	0.54	1.82	0.05
p41	0.72	0.62	0.64	0.66	0.66	4.12	0.07
p42	0.66	0.65	0.6	0.67	0.65	2.47	0.06
p43	0.47	0.54	0.59	0.62	0.56	9.42	0.12
p44	0.61	0.61	0.73	0.78	0.68	16.89**	0.16
p45	0.62	0.47	0.59	0.7	0.60	15.26**	0.15
p46	0.44	0.36	0.49	0.44	0.43	4.52	0.08
p47	0.73	0.74	0.83	0.8	0.78	6.86	0.10
p54	0.79	0.78	0.8	0.83	0.80	1.65	0.05
p55	0.60	0.65	0.63	0.72	0.65	5.97	0.09
p56	0.74	0.83	0.8	0.67	0.76	12.29**	0.13
p57	0.60	0.68	0.75	0.82	0.71	23.26**	0.18
p58	0.69	0.76	0.83	0.83	0.78	14.77**	0.15
p59	0.59	0.65	0.57	0.62	0.61	2.26	0.05
p60	0.66	0.69	0.78	0.84	0.74	17.52**	0.16
p61	0.55	0.57	0.60	0.62	0.59	2.16	0.05
p62	0.40	0.38	0.42	0.43	0.41	0.77	0.03
p63	0.69	0.65	0.75	0.83	0.73	14.01**	0.14
Máx.	0.79	0.83	0.83	0.84	0.80		0.18
M.	0.61	0.61	0.64	0.67	0.63		0.10
Mín.	0.4	0.36	0.42	0.43	0.41		0.03

[a]: El nivel α ajustado por Bonferroni (α/n), en que *n* es el número elementos para comparar es 0.05 / 4 = 0.012; **: p < 0.01

Conceptos Cuantitativos. El análisis intra-año, solo en el año 2004 se detectaron diferencias alrededor del nivel moderado en 5 ítems, y estas favorecieron sistemáticamente al colegio M.I. (Tabla 13) Esto indica que estos ítems tendieron a ser

más fáciles para los niños de este colegio; sin embargo, esta diferencia no se repitió confiablemente en las comparaciones intra-año siguientes ya que las diferencias fueron menores al nivel moderado, llegando incluso a diferencias cero. En el resto de las comparaciones intra-año, la magnitud del efecto tendieron a acercarse a cero, y estas tuvieron signo positivo. Esto señala que en todas las comparaciones hubo relativamente más respuestas correctas en los niños del colegio M.I y no en el colegio S.M., pero la mayoría de estas diferencias fueron apenas perceptibles y de poca importancia práctica. El ítem p76 no existió en el año 2004, y por este motivo no se calcularon el cálculo del χ^2 y ni *d* de Cohen.

Se debe hacer notar que mientras más grandes fueron las diferencias intra-año, menos se repetían en las otras comparaciones intra-año. La presencia de este patrón parece sugerir que las diferencias de mayor tamaño podrían ser eventos únicos que no se replicaron en las demás comparaciones.

En el siguiente nivel de análisis, se examinaron las posibles diferencias entre-años (Tabla 14). Aquí se halló que el 40% de los ítems mostró variabilidad estadística ($p < 0.001$), y con magnitudes de efecto entre 0.17 y 0.32. Esta variabilidad se debió a efectos sistemáticos más allá del error de muestreo. Se observa que en el año 2004 hubo más proporción de respuestas correctas que los demás años; la dificultad, entonces, no se mantuvo constante en estos ítems y tuvo como eje principal la mayor facilidad para los niños del año 2004. Se debe hacer notar que ninguno de los ítems con las mayores discrepancias entre los años se movió hacia la clasificación *fácil* o *difícil*; por lo tanto. Ninguna de estas diferencias, sin embargo, supera un V_{cramer} de 0.25; es decir, no superan el 6% de varianza explicada por estas diferencias.

Tabla 13
Escala: Hab. Cuantitativas: Diferencia arcsine en la dificultad de los ítems, entre los colegios en cada año muestrado

	2004			2006			2007		
	MI	SM	d[a]	MI	SM	d[a]	MI	SM	d[a]
p64	0.67	0.54	0.27	0.44	0.40	0.08	0.52	0.37	0.30
p65	0.72	0.65	0.15	0.72	0.73	-0.02	0.80	0.72	0.19
p66	0.63	0.73	-0.21	0.62	0.62	0.00	0.59	0.61	-0.04
p67	0.8	0.84	-0.10	0.80	0.89	-0.25	0.89	0.86	0.09
p68	0.59	0.7	-0.23	0.62	0.54	0.16	0.60	0.56	0.08
p69	0.68	0.48	0.41	0.72	0.48	0.50	0.64	0.61	0.06
p70	0.73	0.57	0.34	0.71	0.70	0.02	0.70	0.68	0.04
p71	0.63	0.54	0.18	0.51	0.56	-0.10	0.60	0.55	0.10
p72	0.81	0.73	0.19	0.83	0.84	-0.03	0.88	0.91	-0.10
p73	0.88	0.67	0.52	0.75	0.77	-0.05	0.88	0.82	0.17
p74	0.83	0.73	0.24	0.83	0.76	0.17	0.90	0.85	0.15
p75	0.86	0.7	0.39	0.77	0.86	-0.23	0.90	0.91	-0.03
p76	--	--	--	0.74	0.68	0.13	0.80	0.55	0.54
p77	0.91	0.77	0.39	0.76	0.78	-0.05	0.82	0.86	-0.11
p78	0.9	0.81	0.26	0.84	0.81	0.08	0.89	0.87	0.06
p79	0.8	0.69	0.25	0.62	0.66	-0.08	0.76	0.68	0.18
p80	0.61	0.52	0.18	0.51	0.56	-0.10	0.72	0.55	0.36
p81	0.52	0.32	0.41	0.63	0.62	0.02	0.74	0.64	0.22
p82	0.6	0.49	0.22	0.58	0.53	0.10	0.38	0.32	0.13
p83	0.6	0.53	0.14	0.45	0.40	0.10	0.45	0.49	-0.08
p84	0.58	0.44	0.28	0.62	0.59	0.06	0.60	0.57	0.06
p85	0.69	0.69	0.00	0.35	0.51	-0.32	0.46	0.34	0.25
p86	0.91	0.77	0.39	0.69	0.71	-0.04	0.70	0.68	0.04
Máx.	0.91	0.81	0.52	0.84	0.86	0.17	0.90	0.91	0.54
M.	0.74	0.62	0.24 [b]	0.66	0.67	0.10 [b]	0.72	0.66	0.15 [b]
Mín.	0.52	0.32	0.14	0.35	0.40	-0.32	0.38	0.32	-0.11

[a]: diferencias estandarizadas basadas en la transformación arcsine de la dificultad de los ítems.
[b]: promedio obtenido de los valores absolutos

Tabla 14

Escala: Hab. Cuantitativas: Dificultad de los ítems para las muestras totales en cada año

	Estimaciones de dificultad					Prueba[a] y magnitud del efecto	
	2004	2005	2006	2007	$\bar{X}_p$	χ^2	V_{cramer}
p64	0.60	0.54	0.42	0.44	0.50	15.41**	0.15
p65	0.68	0.75	0.73	0.76	0.73	3.30	0.07
p66	0.68	0.62	0.62	0.6	0.63	2.79	0.06
p67	0.82	0.84	0.84	0.87	0.84	1.67	0.05
p68	0.65	0.61	0.58	0.58	0.61	2.53	0.06
p69	0.58	0.61	0.6	0.62	0.60	0.63	0.03
p70	0.65	0.6	0.7	0.69	0.66	3.01	0.07
p71	0.58	0.63	0.53	0.57	0.58	2.86	0.06
p72	0.77	0.79	0.83	0.89	0.82	9.57	0.12
p73	0.77	0.79	0.76	0.84	0.79	3.95	0.07
p74	0.78	0.75	0.8	0.87	0.80	7.30	0.10
p75	0.78	0.79	0.82	0.9	0.82	10.04	0.12
p76	--	0.63	0.71	0.67	0.67	2.04	0.05
p77	0.84	0.64	0.77	0.84	0.77	20.07**	0.18
p78	0.85	0.79	0.82	0.88	0.84	4.65	0.08
p79	0.75	0.87	0.64	0.72	0.75	18.81**	0.17
p80	0.57	0.76	0.53	0.63	0.62	16.56**	0.16
p81	0.42	0.6	0.63	0.69	0.59	30.24**	0.21
p82	0.55	0.67	0.55	0.35	0.53	30.32**	0.22
p83	0.57	0.63	0.42	0.47	0.52	16.01**	0.16
p84	0.51	0.52	0.61	0.59	0.56	5.11	0.08
p85	0.69	0.53	0.43	0.4	0.51	37.51**	0.24
p86	0.84	0.39	0.7	0.69	0.66	64.63**	0.32
Máx.	0.85	0.87	0.83	0.90	0.84		0.32
M.	0.68	0.67	0.66	0.69	0.67		0.12
Mín.	0.42	0.39	0.42	0.35	0.50		0.03

[a]: El nivel α ajustado por Bonferroni (α/n), en que n es el número elementos para comparar es 0.05 / 4 = 0.012; **: $p < 0.01$

Habilidades de Vocabulario/Conceptos. El análisis intra-año de estos ítems indica que, la diferencia promedio de las diferencias estandarizadas en cada comparación fue alrededor de 0.13 (Tabla 15); este bajo nivel sugiere que las discrepancias son en general pequeñas. Justamente, el examen de cada ítem indica que estas diferencias apenas superan 0.30, excepto dos ítems en el 2006 (ítem 89 e ítem 92) y cuatro ítems en el 2007 (ítem 91, 92, 94 y 95). Además se puede observar que las diferencias, en general, favorecen al colegio MI, pues hay la tendencia de que los niños muestreados en este colegio responden con más aciertos que los niños del colegio SM. Los resultados de este análisis sugieren que, dentro de cada periodo, los ítems son menos homogéneos en la dificultad comparándolos con las demás subescalas; y esta variabilidad indica que el contenido de los ítems es relativamente sensible a los conocimientos entre la procedencia institucional de los niños evaluados. Aún con estas diferencias, la magnitud no es grande, y toda la muestra dentro de cada año puede fusionarse para el análisis entre-años. Esta relativa homogeneidad intra-año, como en las subescalas anteriores puede considerarse como la característica central de los ítems. Debe señalarse que la dificultad media en los colegios SM y MI a través de los años fue aproximadamente > 0.65, aumentado hacia los años 2006 y 2007. En general, los ítems de esta área aparecen como levemente más fáciles que las demás subescalas. El rango de dificultad y los valores medios no parecen adecuarse a los criterios establecidos para seleccionar ítems, por lo tanto estos resultados no dan un favorable respaldo a la construcción de esta subescala en relación a las otras.

Tabla 15
Escala: Hab. Vocabulario/Conceptos: Diferencia arcsine en la dificultad de los ítems, entre los colegios en cada año muestrado

	2004			2006			2007		
	MI	SM	d[a]	MI	SM	d[a]	MI	SM	d[a]
p87	0.8	0.83	-0.08	0.82	0.90	-0.23	0.92	0.87	0.16
p88	0.63	0.61	0.04	0.80	0.68	0.28	0.71	0.62	0.19
p89	0.66	0.61	0.10	0.77	0.63	0.31	0.72	0.74	-0.05
p90	0.75	0.67	0.18	0.80	0.84	-0.10	0.84	0.89	-0.15
p91	0.61	0.65	-0.08	0.80	0.85	-0.13	0.88	0.74	0.36
p92	0.49	0.53	-0.08	0.76	0.59	0.37	0.74	0.49	0.52
p93	0.69	0.61	0.17	0.68	0.59	0.19	0.62	0.63	-0.02
p94	0.71	0.69	0.04	0.78	0.77	0.02	0.89	0.76	0.35
p95	0.66	0.61	0.10	0.84	0.77	0.18	0.85	0.70	0.36
p96	--	--	--	0.70	0.56	0.29	0.62	0.48	0.28
p97	0.74	0.52	0.46	0.45	0.44	0.02	0.49	0.48	0.02
p98	0.6	0.6	0.00	0.64	0.54	0.20	0.68	0.68	0.00
p99	--	--	--	0.76	0.83	-0.17	0.82	0.86	-0.11
p100	0.83	0.75	0.20	0.76	0.70	0.14	0.68	0.77	-0.20
p101	--	--	--	0.51	0.37	0.28	0.46	0.47	-0.02
p102	--	--	--	0.94	0.87	0.24	0.92	0.93	-0.04
p103	--	--	--	0.73	0.70	0.07	0.74	0.79	-0.12
p104	0.25	0.35	-0.22	0.84	0.87	-0.09	0.88	0.79	0.24
p105	0.8	0.77	0.07	0.72	0.74	-0.05	0.84	0.79	0.13
p106	--	--	--	0.73	0.83	-0.24	0.86	0.84	0.06
Máx.	0.83	0.83	0.46	0.94	0.90	0.37	0.92	0.93	0.52
M.	0.66	0.63	0.09[b]	0.74	0.70	0.18[b]	0.76	0.72	0.17[b]
Mín.	0.25	0.35	-0.22	0.45	0.37	-0.24	0.46	0.47	-0.20

[a]: diferencias estandarizadas basadas en la transformación arcsine de la dificultad de los ítems.
[b]: promedio obtenido de los valores absolutos

Tabla 16
Escala: Hab. Vocabulario/Conceptos: Dificultad de los ítems para las muestras totales en cada año

	Estimaciones de dificultad					Prueba[a] y magnitud del efecto	
	2004	2005	2006	2007	$\bar{X}_r$	χ^2	V_{cramer}
p87	0.81	0.64	0.86	0.90	0.80	32.81**	0.22
p88	0.62	0.86	0.74	0.66	0.72	21.62**	0.18
p89	0.63	0.62	0.70	0.73	0.67	6.04	0.09
p90	0.71	0.68	0.82	0.86	0.77	19.22**	0.17
p91	0.63	0.83	0.82	0.80	0.77	25.62**	0.19
p92	0.51	0.77	0.67	0.61	0.64	22.02**	0.18
p93	0.65	0.65	0.64	0.63	0.64	0.19	0.01
p94	0.70	0.61	0.78	0.82	0.73	18.03**	0.16
p95	0.63	0.65	0.81	0.77	0.72	19.82**	0.17
p96	--	0.69	0.63	0.55	0.62	5.69	0.11
p97	0.63	0.57	0.44	0.49	0.53	15.42**	0.15
p98	0.60	0.45	0.59	0.68	0.58	14.34**	0.14
p99	--	0.64	0.79	0.84	0.76	15.38**	0.18
p100	0.79	0.73	0.73	0.72	0.74	2.91	0.06
p101	--	0.75	0.44	0.47	0.55	29.14**	0.25
p102	--	0.50	0.90	0.93	0.78	95.2**	0.45
p103	--	0.91	0.72	0.77	0.80	14.65**	0.17
p104	0.30	0.76	0.86	0.83	0.69	171.58**	0.51
p105	0.79	0.89	0.73	0.81	0.81	11.05	0.13
p106	--	0.78	0.78	0.85	0.80	3.32	0.08
Máx.	0.79	0.91	0.90	0.93	0.81		0.51
M.	0.63	0.70	0.72	0.74	0.71		0.18
Mín.	0.30	0.45	0.44	0.47	0.53		0.01

[a]: El nivel α ajustado por Bonferroni (α/n), en que n es el número elementos para comparar es 0.05 / 4 = 0.012; **: $p < 0.01$

Discriminación del ítem

Discriminación del ítem en Conocimiento de letras y palabras. En la comparación intra-año de los ítems de esta subescala, la magnitud de la discriminación de los ítems ha tendido a ser similar entre los colegios M.I y S.M (Tabla 17). Este patrón de similaridad es constante en los ítems de esta subescala en los periodos muestreados, lo que se confirmó con la prueba z de comparación de correlaciones independientes; las pruebas z aplicadas estuvieron alrededor de 0.70, y por lo tanto las diferencias no son consideradas estadísticamente significativas. Únicamente los ítems 15 y 19 relevaron diferencias entre los colegios (periodo 2006 y 2007) más allá de criterio $z = \pm 1.96$ (*correspondiente a* $p < 0.05$), pero luego del ajuste Bonferroni, las diferencias halladas en ambos ítems podrían no ocurrir en la población. Luego de unir las muestras de ambos colegios, el análisis de la discriminación entre los periodos (Tabla 18) también indicó que los coeficientes ítem-test no varían ni sistemáticamente en cada año muestreado, ni en gran magnitud, pues la V_{Cramer} aplicada en cada ítem muestra una pequeña magnitud. En general, los resultados muestran que lea discriminación de los ítems en esta subescala se ha mantenido en aceptables mínimos niveles.

Tabla 17
Subescala Letras y Palabras: Diferencia de la discriminación de los ítems en los colegios, entre cada año muestreado

	2004			2006			2007		
	MI	SM	Z [a]	MI	SM	Z [a]	MI	SM	Z [a]
p8	0.22	0.38	-1.19	0.28	0.30	-0.15	0.23	0.20	0.20
p9	0.36	0.43	-0.56	0.35	0.38	-0.23	0.38	0.17	1.45
p10	0.50	0.52	-0.18	0.43	0.40	0.24	0.26	0.37	-0.78
p11	0.46	0.57	-1.02	0.49	0.36	1.07	0.30	0.20	0.68
p12	0.45	0.52	-0.62	0.29	0.22	0.50	0.49	0.43	0.48
p13	0.55	0.56	-0.10	0.49	0.46	0.26	0.35	0.36	-0.07
p14	0.61	0.44	1.60	0.54	0.56	-0.19	0.49	0.44	0.40
p15	0.32	0.40	-0.62	0.46	0.48	-0.17	0.41	0.11	2.06
p16	0.49	0.49	0.00	0.61	0.51	0.98	0.36	0.39	-0.22
p17	0.52	0.55	-0.28	0.45	0.40	0.41	0.34	0.32	0.14
p18	0.62	0.69	-0.83	0.67	0.62	0.58	0.57	0.64	-0.70
p19	0.44	0.62	-1.71	0.69	0.43	2.61	0.53	0.59	-0.55
p48	0.33	0.28	0.37	0.36	0.26	0.75	0.38	0.40	-0.15
p49	0.25	0.26	-0.07	0.28	0.37	-0.68	0.58	0.40	1.51
p50	0.46	0.43	0.25	0.47	0.50	-0.26	0.50	0.35	1.17
p51	0.52	0.66	-1.46	0.64	0.65	-0.12	0.63	0.66	-0.33
p52	0.55	0.67	-1.30	0.71	0.53	2.00	0.62	0.50	1.11
p53	0.47	0.49	-0.17	0.74	0.66	1.061	0.58	0.71	-1.42
Máx.	0.62	0.69	1.71	0.74	0.66	2.61	0.63	0.71	2.06
M.	0.45	0.50	0.69	0.50	0.45	0.68	0.44	0.40	0.75
Mín.	0.22	0.26	0.00	0.28	0.22	0.12	0.23	0.11	0.07

[a]: prueba Z para correlaciones independientes, con corrección Bonferroni, $\alpha = 0.002$.
[b]: en cada ítem, la media y el valor mínimo de la prueba z se obtuvieron con los valores absolutos de las correlaciones

Tabla 18

Subescala Letras y Palabras: Diferencia de la discriminación de los ítems en la muestra total, a través de los años

	2004	2005	2006	2007	$\bar{X}_r$	$\chi^{2,\,a}$	V_{cramer}
p8	0.37	0.34	0.29	0.22	0.31	2.58	0.06
p9	0.44	0.47	0.37	0.27	0.39	4.73	0.09
p10	0.53	0.44	0.42	0.32	0.43	5.89	0.10
p11	0.52	0.46	0.43	0.25	0.42	9.43	0.12
p12	0.50	0.51	0.26	0.46	0.43	9.72	0.12
p13	0.59	0.37	0.48	0.36	0.45	9.70	0.12
p14	0.54	0.51	0.55	0.47	0.52	1.21	0.04
p15	0.40	0.48	0.47	0.26	0.40	6.47	0.10
p16	0.51	0.44	0.56	0.38	0.47	5.25	0.09
p17	0.56	0.35	0.42	0.33	0.42	8.76	0.12
p18	0.66	0.56	0.64	0.61	0.62	1.92	0.05
p19	0.54	0.61	0.47	0.56	0.55	2.90	0.07
p48	0.32	0.48	0.30	0.39	0.37	3.64	0.07
p49	0.29	0.42	0.33	0.48	0.38	5.20	0.09
p50	0.45	0.48	0.49	0.41	0.46	0.99	0.04
p51	0.59	0.70	0.65	0.64	0.65	2.50	0.06
p52	0.63	0.68	0.62	0.55	0.62	3.02	0.07
p53	0.51	0.58	0.70	0.64	0.61	9.17	0.12
Máx	0.66	0.70	0.70	0.64	0.65		0.12
M	0.50	0.49	0.47	0.42	0.47		0.09
Mín.	0.29	0.34	0.26	0.22	0.31		0.04

[a]: Prueba χ^2 con corrección Bonferroni: α= 0.012

Discriminación del ítem en Habilidades Fonológicas. Algunos ítems mostraron diferencias más allá del error de muestreo luego de la corrección Bonferroni al alfa crítico (0.05), como los ítems 20, 21 y 34 del periodo 2006 (Tabla 19), pero eso no volvió a ocurrir en las demás comparaciones intra-año; esto significa que en los demás ítems del resto de las comparaciones intra-año no hubo evidencias para rechazar la hipótesis nula de igualdad de correlaciones en la población. La magnitud promedio de la correlación ítem test fue > 0.30; y el patrón de magnitud ligeramente elevadas correspondió al colegio MI frente al colegio SM en los periodos 2006 y 2007. Sin discrepancias consistentes ni de gran magnitud, se puede asumir que el efecto del colegio de procedencia es trivial, y pasar al siguiente análisis entre-años que se presenta en la Tabla 20. Los resultados muestran homogeneidad de los valores promedio, mínimo y máximo en los coeficientes de discriminación en cada año; y consecuentemente, las diferencias entre ellas fueron consideradas como error de muestreo. De acuerdo a la Tabla 20, las pequeñas discrepancias entre los coeficientes de discriminación en cada años explicaron menos del 0.05% de la varianza ($\overline{X}\ V_{Cramer} < 0.07$).

Tabla 19
Subescala Hab. Fonológicas: Diferencia de la discriminación de los ítems en los colegios, entre cada año muestreado

	2004			2006			2007		
	MI	SM	z [a]	Mi	SM	z [a]	Mi	SM	z [a]
p20	0.34	0.54	-1.69	0.66	0.32	3.10**	0.47	0.42	0.40
p21	0.38	0.59	-1.88	0.68	0.30	3.50**	0.51	0.46	0.41
p22	0.29	0.57	-2.36	0.60	0.39	1.89	0.49	0.47	0.16
p23	0.38	0.19	1.40	0.49	0.27	1.74	0.37	0.19	1.24
p24	0.37	0.32	0.38	0.32	0.16	1.15	0.23	0.05	1.17
p25	0.45	0.31	1.11	0.45	0.47	-0.17	0.42	0.36	0.45
p26	0.31	0.27	0.30	0.29	0.20	0.64	0.37	0.37	0.00
p27	0.31	0.41	-0.78	0.39	0.23	1.19	0.42	0.13	2.01
p28	0.34	0.32	0.15	0.40	0.24	1.20	0.21	0.43	-1.56
p29	0.35	0.52	-1.43	0.32	0.34	-0.15	0.52	0.47	0.42
p30	0.28	0.44	-1.25	0.36	0.28	0.60	0.32	0.17	1.01
p31	0.43	0.41	0.16	0.37	0.40	-0.24	0.38	0.26	0.85
p32	0.28	0.54	-2.14	0.42	0.32	0.78	0.37	0.25	0.84
p33	0.49	0.45	0.35	0.53	0.40	1.12	0.46	0.43	0.24
p34	0.48	0.42	0.51	0.59	0.35	2.10**	0.47	0.50	-0.25
p35	0.41	0.56	-1.33	0.33	0.36	-0.23	0.56	0.43	1.10
p36	0.36	0.39	-0.24	0.51	0.49	0.18	0.55	0.50	0.44
Máx.	0.49	0.59	2.36	0.68	0.49	3.50	0.56	0.50	2.01
M [b]	0.37	0.43	1.03	0.45	0.32	1.18	0.42	0.35	0.74
Mín.	0.28	0.19	0.15	0.29	0.16	0.15	0.21	0.05	0.00

[a]: prueba Z para correlaciones independientes, con corrección Bonferroni, $\alpha = 0.0022$
[b]: en cada ítem, la media y el valor mínimo de la prueba z se obtuvieron con los valores absolutos de las correlaciones

Tabla 20
Subescala Hab. Fonológicas: Discriminación de los ítems en la muestra total a través de los años

	2004	2005	2006	2007	$\bar{X}_r$	χ^2	V_{cramer}
p20	0.45	0.53	0.50	0.44	0.48	1.27	0.04
p21	0.51	0.53	0.50	0.48	0.51	0.31	0.02
p22	0.46	0.51	0.50	0.48	0.49	0.39	0.02
p23	0.29	0.32	0.39	0.28	0.32	1.70	0.05
p24	0.37	0.36	0.24	0.10	0.27	8.63	0.11
p25	0.37	0.51	0.45	0.39	0.43	2.49	0.06
p26	0.31	0.35	0.25	0.36	0.32	1.50	0.04
p27	0.38	0.30	0.32	0.28	0.32	1.22	0.04
p28	0.33	0.49	0.33	0.33	0.37	3.25	0.07
p29	0.45	0.51	0.32	0.50	0.45	5.49	0.09
p30	0.38	0.37	0.32	0.25	0.33	2.10	0.05
p31	0.43	0.47	0.38	0.32	0.40	2.49	0.06
p32	0.44	0.51	0.37	0.30	0.41	4.81	0.08
p33	0.47	0.44	0.47	0.44	0.46	0.22	0.01
p34	0.40	0.37	0.48	0.49	0.44	2.30	0.05
p35	0.51	0.38	0.34	0.49	0.43	5.30	0.09
p36	0.40	0.35	0.50	0.52	0.44	4.30	0.08
Máx.	0.51	0.53	0.50	0.52	0.51		0.11
M	0.41	0.43	0.39	0.38	0.41		0.06
Mín.	0.29	0.30	0.24	0.10	0.29		0.01

Prueba χ^2 con corrección Bonferroni: $\alpha = 0.012$

Discriminación del ítem en Habilidades de Percepción Visual. Excepto el ítem 58 (Tabla 21), el resto no mostró consistentes diferencias entre los colegios. Algunos ítems mostraron un patrón de diferencias que favoreció al colegio SM (años 2006 y 2007), pero estas no fueron de gran magnitud. En el año 2007, se detectaron diferencias en estadísticamente significativas en los ítems 47, 54, 58 y 60 entre los colegios, y favoreciendo al colegio MI; en el año 2004 también se detectó una diferencia estadística, pero esta vez favoreciendo al colegio SM. El ítem 47 tendió a ser bajo, incluso mostrando una correlación de 0 (año 2007), y podría sugerirse su remoción.

Tabla 21

Subescala Hab. Percepción Visual: Diferencia de la discriminación de los ítems en los colegios, entre cada año muestreado

	2004			2006			2007		
	Mi	SM	Z [a]	Mi	SM	Z [a]	Mi	SM	Z [a]
p37	0.30	0.13	1.20	0.28	0.20	0.59	0.45	0.29	1.26
p38	0.25	0.34	-0.67	0.43	0.26	1.34	0.34	0.42	-0.57
p39	0.32	0.15	1.23	0.25	0.34	-0.66	0.41	0.37	0.26
p40	0.40	0.30	0.74	0.51	0.33	1.49	0.41	0.40	0.10
p41	0.29	0.31	-0.11	0.54	0.37	1.40	0.43	0.43	0.02
p42	0.36	0.24	0.90	0.36	0.09	1.94	0.37	0.35	0.15
p43	0.30	0.33	-0.21	0.35	0.37	-0.12	0.40	0.31	0.70
p44	0.30	0.30	0.02	0.43	0.37	0.53	0.36	0.28	0.60
p45	0.32	0.33	-0.08	0.39	0.29	0.77	0.48	0.44	0.37
p46	0.00	0.22	-1.53	0.34	0.33	0.08	0.44	0.25	1.48
p47	0.28	0.19	0.64	0.43	0.23	1.54	0.43	-0.04	3.36**
p54	0.40	0.44	-0.32	0.54	0.32	1.89	0.64	0.19	3.78**
p55	0.43	0.40	0.24	0.45	0.40	0.37	0.52	0.41	1.00
p56	0.40	0.47	-0.58	0.51	0.44	0.62	0.49	0.27	1.79
p57	0.29	0.40	-0.84	0.42	0.49	-0.56	0.61	0.30	2.68
p58	0.26	0.52	-2.10**	0.54	0.41	1.09	0.61	0.24	3.21**
p59	0.25	0.50	-1.98	0.48	0.45	0.31	0.50	0.23	2.13
p60	0.47	0.54	-0.58	0.59	0.52	0.71	0.72	0.21	4.61**
p61	0.39	0.53	-1.25	0.46	0.55	-0.84	0.53	0.22	2.49
p62	0.22	0.07	1.01	0.23	0.28	-0.40	0.26	0.04	1.53
p63	0.35	0.49	-1.13	0.52	0.51	0.11	0.66	0.41	2.46
Máx.	0.47	0.54	2.10	0.59	0.55	1.94	0.72	0.44	4.61
M [b]	0.31	0.34	0.83	0.43	0.36	0.83	0.48	0.29	1.65
Mín.	0.00	0.07	0.02	0.23	0.09	0.08	0.26	-0.04	0.02

[a]: prueba Z para correlaciones independientes, con corrección Bonferroni: $\alpha = 0.0023$.

[b]: en cada ítem, la media y el valor mínimo de la prueba z se obtuvieron con los valores absolutos de las correlaciones

Tabla 22

Subescala Hab. Percepción Visual: Discriminación de los ítems en la muestra total a través de los años

	2004	2005	2006	2007	$\bar{X}_r$	$\chi^{2,\,a}$	V_{cramer}
p37	0.22	0.32	0.24	0.37	0.29	2.96	0.06
p38	0.29	0.39	0.35	0.37	0.35	1.12	0.04
p39	0.23	0.32	0.29	0.39	0.31	2.83	0.06
p40	0.35	0.32	0.43	0.40	0.37	1.46	0.04
p41	0.30	0.47	0.46	0.42	0.41	4.25	0.08
p42	0.29	0.40	0.24	0.35	0.32	2.58	0.06
p43	0.32	0.42	0.36	0.35	0.36	0.91	0.03
p44	0.30	0.49	0.40	0.31	0.38	4.38	0.08
p45	0.32	0.51	0.34	0.46	0.41	6.35	0.09
p46	0.10	0.35	0.33	0.36	0.28	9.02	0.11
p47	0.24	0.37	0.34	0.23	0.29	2.62	0.06
p54	0.42	0.46	0.44	0.45	0.44	0.20	0.01
p55	0.41	0.54	0.42	0.46	0.46	2.21	0.05
p56	0.42	0.57	0.46	0.38	0.46	4.21	0.08
p57	0.34	0.52	0.45	0.48	0.45	4.17	0.08
p58	0.38	0.58	0.48	0.45	0.47	4.74	0.08
p59	0.38	0.56	0.47	0.38	0.45	4.76	0.08
p60	0.51	0.60	0.55	0.51	0.54	1.45	0.04
p61	0.46	0.55	0.50	0.38	0.47	3.51	0.07
p62	0.14	0.39	0.25	0.16	0.24	5.78	0.09
p63	0.42	0.51	0.52	0.55	0.50	2.85	0.06
Máx.	0.51	0.60	0.55	0.55	0.54		0.11
M	0.33	0.46	0.40	0.39	0.39		0.06
Mín.	0.10	0.32	0.24	0.16	0.24		0.01

[a] Prueba χ^2 con corrección Bonferroni: $\alpha = 0.012$

Discriminación del ítem en Habilidades Cuantitativas. Los resultados del análisis intra-año se muestran en la Tabla 23. La discriminación promedio de los ítems en esta escala fue > 0.35, indicando que existe la tendencia que esta subescala posea ítems con buena capacidad discriminativa. Sin embargo, hubo algunas diferencias más allá del error de muestreo. El ítem 72 se diferenció en su capacidad discriminativa en el periodo 2006, pero debido a la corrección Bonferroni, la prueba Z no fue estadísticamente significativa. Al mismo, el ítem 67 fue sistemáticamente bajo en la muestra del colegio SM en todos los años, y puede requerir ser revisado o eliminado. Se detectó una diferencia estadística en el ítem 82, en que el coeficiente ítem-test fue mayor en el colegio SM. Una magnitud similar o más elevada no se repitió en los otros años, por lo tanto se podría concluir en que esta variación fue idiosincrásica a este periodo. Por otro lado, el resto de los ítems no mostraron elevadas diferencias ni diferencias estadísticamente significativas; tampoco hubo un claro patrón que pudiera informar que la discriminación de los ítems en general podría favorecer a uno de los colegios.

Por otro lado, en el análisis entre-años (Tabla 24), se halló que los ítems 74, 79, 81, 82 y 83 difieren en la magnitud de los coeficientes. Pero se puede observar que las diferencias están asociadas a un coeficiente extremo (outlier) en cada uno de estos ítems, y que cuya elevada magnitud es algo frecuente en el año 2006. Las diferencias en estos coeficientes representan menos de $V_{cramer} < 0.19$, que sugiere una magnitud pequeña. Las variaciones entre los años están alrededor de una media de 0.40; y exceptuando algunas r_{itc} debajo de 0.20, el resto se mantiene en niveles recomendables de poder discriminativo.

Tabla 23
Subescala Hab. Cuantitativas: Diferencia de la discriminación de los ítems en los colegios, entre cada año muestreado

	2004			2006			2007		
	MI	SM	Z[a]	MI	SM	Z[a]	MI	SM	Z[a]
p64	0.47	0.53	-0.56	0.37	0.35	0.18	0.37	0.34	0.21
p65	0.61	0.56	0.53	0.56	0.55	0.09	0.57	0.34	2.01
p66	0.39	0.27	0.88	0.45	0.33	0.94	0.31	0.37	-0.45
p67	0.39	0.22	1.29	0.43	0.17	1.91	0.11	0.00	0.72
p68	0.32	0.13	1.34	0.38	0.34	0.29	0.33	0.23	0.71
p69	0.36	0.52	-1.30	0.47	0.45	0.15	0.34	0.44	-0.79
p70	0.39	0.28	0.85	0.46	0.23	1.72	0.45	0.40	0.45
p71	0.35	0.52	-1.40	0.41	0.38	0.18	0.37	0.33	0.30
p72	0.45	0.60	-1.44	0.57	0.32	2.10	0.40	0.35	0.34
p73	0.37	0.65	-2.57	0.64	0.62	0.13	0.55	0.36	1.63
p74	0.50	0.58	-0.82	0.55	0.62	-0.72	0.24	0.39	-1.07
p75	0.46	0.55	-0.82	0.53	0.45	0.68	0.49	0.37	0.96
p76				0.74	0.53	2.40	0.38	0.43	-0.40
p77	0.38	0.50	-1.06	0.58	0.47	1.01	0.43	0.50	-0.54
p78	0.43	0.49	-0.49	0.48	0.39	0.76	0.35	0.43	-0.63
p79	0.44	0.43	0.06	0.64	0.72	-1.00	0.47	0.46	0.11
p80	0.56	0.47	0.82	0.67	0.72	-0.69	0.42	0.57	-1.38
p81	0.44	0.48	-0.32	0.80	0.67	1.97	0.56	0.48	0.73
p82	0.37	0.36	0.086	0.61	0.28	2.86**	0.29	0.22	0.52
p83	0.24	0.29	-0.379	0.43	0.45	-0.09	0.46	0.27	1.45
p84	0.37	0.23	1.034	0.37	0.23	1.05	0.23	0.23	0.04
p85	0.36	0.49	-1.001	0.28	0.38	-0.75	0.29	0.16	0.90
p86	0.38	0.50	-1.058	0.22	0.22	0.00	0.31	0.06	1.80
Máx.	0.61	0.65	2.57	0.80	0.72	2.86	0.57	0.57	2.01
M[b]	0.41	0.44	0.91	0.51	0.43	0.88	0.38	0.34	0.81
Mín.	0.24	0.13	0.06	0.22	0.17	0.00	0.11	0.00	0.04

[a]: prueba Z para correlaciones independientes, con corrección Bonferroni $\alpha = 0.0022$
[b]: en cada ítem, la media y el valor mínimo de la prueba z se obtuvieron con los valores absolutos de las correlaciones

Tabla 24
Subescala Hab. Cuantitativas: Discriminación de los ítems en la muestra total a través de los años

	2004	2005	2006	2007	$\bar{X}_r$	$\chi^{2,\,a}$	V_{cramer}
p64	0.51	0.43	0.36	0.36	0.42	4.22	0.08
p65	0.58	0.45	0.55	0.45	0.51	3.95	0.07
p66	0.29	0.35	0.39	0.34	0.34	1.20	0.04
p67	0.28	0.31	0.31	0.05	0.24	8.21	0.11
p68	0.19	0.33	0.36	0.28	0.29	3.44	0.07
p69	0.47	0.50	0.44	0.39	0.45	3.85	0.07
p70	0.35	0.20	0.35	0.42	0.33	3.84	0.07
p71	0.45	0.40	0.39	0.35	0.40	1.27	0.04
p72	0.54	0.54	0.45	0.36	0.47	5.59	0.09
p73	0.56	0.43	0.63	0.45	0.52	8.10	0.11
p74	0.56	0.45	0.58	0.33	0.48	11.05**	0.13
p75	0.53	0.57	0.49	0.42	0.50	3.04	0.06
p76	--	0.69	0.63	0.42	0.58	12.29	0.16
p77	0.47	0.59	0.53	0.45	0.51	2.98	0.06
p78	0.48	0.55	0.43	0.39	0.46	3.08	0.06
p79	0.45	0.53	0.67	0.47	0.53	11.98**	0.13
p80	0.52	0.58	0.69	0.51	0.57	9.34	0.11
p81	0.48	0.52	0.74	0.52	0.56	20.85**	0.17
p82	0.38	0.65	0.45	0.26	0.44	17.24**	0.16
p83	0.28	0.68	0.44	0.35	0.44	21.11**	0.18
p84	0.31	0.49	0.30	0.23	0.33	6.02	0.09
p85	0.42	0.44	0.32	0.24	0.35	5.04	0.08
p86	0.47	0.34	0.22	0.18	0.30	11.67	0.13
Máx.	0.58	0.69	0.74	0.52	0.58		0.18
M	0.44	0.48	0.47	0.36	0.44		0.10
Mín.	0.19	0.20	0.22	0.05	0.24		0.04

[a]: Prueba χ^2 con corrección Bonferroni: $\alpha = 0.012$

Discriminación del ítem en Vocabulario/Conceptos. Los resultados del análisis intra-años (Tabla 25) indican algunos patrones que podrían poner en cuestión la integridad de esta subescala. Primero, en dos periodos de evaluación, hubo una variada magnitud mayor entre los colegios. Es decir, este patrón no fue consistente, pues en el año 2004, la magnitud promedio de las r_{itc} fue similar entre los colegios, en el año 2006 la tendencia favoreció al colegio MI, y en el año 2007 la tendencia promedio favoreció el

colegio SM. También, un ítem fue favorable para un colegio en dos periodos (ítem 89), y otros dos ítems favorecieron a uno de los colegios (ítem 94 e ítem 105). Estas diferencias, aunque son estadísticamente significativas luego de la corrección Bonferroni, son relativamente pequeñas. Pero aun cuando el resto de contrastes no alcanzan significancia estadística, parece identificarse que esta subescala es sensible a variaciones entre los niños más que las otras subescalas.

Tabla 25
Subescala Hab. Vocabulario/Conceptos: Diferencia de la discriminación de los ítems en los colegios, entre cada año muestreado

	2004			2006			2007		
	MI	SM	Z[a]	MI	SM	Z[a]	MI	SM	Z[a]
p87	0.12	0.16	-0.28	0.43	0.18	1.85	0.03	0.40	-2.70
p88	0.42	0.53	-0.95	0.38	0.34	0.37	0.20	0.43	-1.71
p89	0.71	0.29	3.97**	0.44	0.03	2.96**	0.18	0.49	-2.40
p90	0.31	0.34	-0.25	0.37	0.24	0.99	0.22	0.26	-0.33
p91	0.58	0.55	0.32	0.38	0.24	1.03	0.41	0.52	-0.97
p92	0.42	0.41	0.08	0.42	0.08	2.52	0.18	0.41	-1.73
p93	0.39	0.39	-0.06	0.18	0.03	1.01	0.21	0.12	0.60
p94	0.49	0.57	-0.74	0.43	0.28	1.20	0.22	0.59	-2.99**
p95	0.71	0.54	1.95	0.54	0.31	1.93	0.43	0.53	-0.82
p96				0.26	0.31	-0.34	0.23	0.45	-1.68
p97	0.41	0.42	-0.06	0.29	0.23	0.43	0.22	0.41	-1.48
p98	0.36	0.35	0.10	0.32	-0.02	2.32	0.32	0.28	0.28
p99				0.32	0.08	1.65	0.06	0.04	0.12
p100	0.19	0.27	-0.58	0.49	0.25	1.83	0.09	0.36	-1.93
p101				0.29	0.18	0.74	0.13	0.31	-1.31
p102				0.29	0.23	0.43	0.27	0.34	-0.55
p103				0.41	0.19	1.63	0.10	0.04	0.42
p104	-0.25	-0.10	-1.02	0.13	0.10	0.18	0.01	0.21	-1.33
p105	0.34	0.40	-0.50	0.45	0.20	1.89	0.00	0.49	-3.55**
p106				0.33	0.13	1.48	0.20	0.09	0.79
Máx.	0.71	0.57	3.97	0.54	0.34	2.96	0.43	0.59	3.55
M[b]	0.37	0.37	0.78	0.36	0.18	1.34	0.19	0.34	1.38
Mín.	-0.25	-0.10	0.06	0.13	-0.02	0.18	0.00	0.04	0.12

[a]: prueba Z para correlaciones independientes, con corrección Bonferroni $\alpha = 0.0022$
[b]: en cada ítem, la media y el valor mínimo de la prueba z se obtuvieron con los valores absolutos de las correlaciones

En la comparación entre años (Tabla 26), apenas se detectaron diferencias en el poder discriminatorio en solo dos ítems (ítem 93 e ítem 104), y estas diferencias fueron sistemáticas al menos para dos periodos de evaluación específico, pues las r_{itc} más baja y alta fueron en los años 2005 y 2005, respectivamente. También parece observarse una tendencia en que las magnitudes bajas se ubican en los años 2006 y 2007. La magnitud del efecto de varias comparaciones están alrededor de 0.10 en 8 ítems, y aunque no alcanzan significancia estadística, puede observarse una relativa heterogeneidad en la discriminación de los ítems de esta subescala.

Tabla 26
Subescala Hab. Vocabulario/Conceptos: Discriminación de los ítems en la muestra total a través de los años

	2004	2005	2006	2007	$\bar{X}_r$	$\chi^{2,\,a}$	V_{cramer}
p87	0.14	0.33	0.32	0.28	0.26	4.39	0.08
p88	0.48	0.37	0.36	0.35	0.39	2.9	0.06
p89	0.49	0.43	0.26	0.36	0.39	7.19	0.10
p90	0.33	0.26	0.31	0.22	0.28	1.47	0.04
p91	0.56	0.44	0.31	0.49	0.45	9.46	0.12
p92	0.41	0.41	0.27	0.34	0.36	2.89	0.06
p93	0.39	0.41	0.12	0.15	0.27	13.00**	0.14
p94	0.53	0.23	0.36	0.47	0.40	10.07	0.12
p95	0.62	0.52	0.43	0.50	0.52	6.72	0.10
p96		0.37	0.29	0.37	0.34	0.88	0.04
p97	0.41	0.44	0.26	0.33	0.36	3.99	0.26
p98	0.35	0.40	0.18	0.29	0.31	4.93	0.08
p99		0.32	0.21	0.04	0.19	5.80	0.05
p100	0.24	0.31	0.38	0.22	0.29	3.40	0.07
p101		0.43	0.25	0.23	0.30	3.72	0.08
p102		0.17	0.26	0.30	0.24	1.22	0.05
p103		0.25	0.31	0.05	0.20	6.66	0.12
p104	-0.18	0.37	0.11	0.15	0.11	23.66**	0.19
p105	0.37	0.18	0.34	0.30	0.30	3.07	0.06
p106		0.31	0.24	0.13	0.23	2.47	0.07
Máx.	0.62	0.52	0.43	0.50	0.52		0.26
M	0.37	0.35	0.28	0.28	0.31		0.09
Mín.	-0.18	0.17	0.11	0.04	0.11		0.04

[a] Prueba χ^2 con corrección Bonferroni: $\alpha = 0.012$

La descripción sumaria del análisis de ítems aparece en la Tabla 27. Visto de este modo, en los periodos muestreados, las escalas muestran parámetros generalmente dentro de los rangos apropiados. La excepción fue la escala Conceptos/Vocabulario, en que la distribución de la discriminación de sus ítems estuvo entre 13% y 26% debajo de 0.20. Este nivel pone en cuestionamiento la integridad de este puntaje.

Tabla 27
Sumario de la dificultad y discriminación

	N ítems	Rangos de Dificultad			Rangos de Discriminación		
		0.40 <	0.40 – 0.89	> 0.90	0.20 <	0.20-0.29	> 0.29
		N (%)	N (%)	N (%)	N (%)	N (%)	N (%)
Hab. Letras y palabras	**18**						
2004		5 (0.3)	13 (72.2)	0 (0)	0 (0.0)	1 (5.6)	17 (94.4)
2005		5 (0.3)	13 (72.2)	0 (0)	0 (0.0)	0 (0)	18 (100)
2006		7 (0.4)	11 (61.1)	0 (0)	0 (0.0)	2 (11.1)	16 (88.9)
2007		4 (0.2)	13 (72.2)	1 (5.6)	0 (0.0)	4 (22.2)	14 (77.8)
Total							
Hab. Fonológicas	**17**						
2004		3 (17.6)	14 (82.4)	0 (0.0)	0 (0.0)	1 (5.9)	16 (94.1)
2005		5 (29.4)	12 (70.6)	0 (0.0)	0 (0.0)	0 (0)	17 (100)
2006		6 (35.3)	11 (64.7)	0 (0.0)	0 (0.0)	2 (11.8)	15 (88.2)
2007		5 (29.4)	12 (70.6)	0 (0.0)	1 (0.1)	3 (17.6)	13 (76.5)
Total							
Percepción Visual	**21**						
2004		0 (0.0)	21 (100)	0 (0.0)	2 (0.1)	7 (30.4)	14 (60.9)
2005		2 (9.5)	19 (90.5)	0 (0.0)	0 (0.0)	3 (12.5)	21 (87.5)
2006		0 (0.0)	21 (100)	0 (0.0)	0 (0.0)	6 (26.1)	17 (73.9)
2007		0 (0.0)	21 (100)	0 (0.0)	1 (0.0)	3 (13)	19 (82.6)
Total							
Conceptos Cuantitativos	**23**						
2004		0(0.0)	22(100)	0(0.0)	1(4.5)	3(13.6)	18(81.8)
2005		1(4.3)	22(95.7)	0(0.0)	0(0.0)	1(4.3)	22(95.7)
2006		0(0.0)	23(100)	0(0.0)	0(0.0)	1(4.3)	22(95.7)
2007		1(4.3)	21(91.3)	1(4.3)	2(8.7)	4(17.4)	17(73.9)
Total							
Conceptos/ Vocabulario	**20**						
2004		4(23.5)	13(76.5)	0(0.0)	3(17.6)	3(17.6)	11(64.7)
2005		0(0.0)	19(95.0)	1(5.0)	3(13)	5(21.7)	15(65.2)
2006		0(0.0)	19(95.0)	1(5.0)	4(17.4)	10(43.5)	9(39.1)
2007		0(0.0)	18(90.0)	2(10.0)	6(26.1)	7(30.4)	10(43.5)

Homogeneidad de los ítems

La homogeneidad de los ítems en la BDPG, muestra correlaciones inter-ítem en las subescalas entre 0.28 y 0.10, y suponen estar midiendo constructos que internamente muestra cierta heterogeneidad debido a las tareas presentadas (Tabla 28).

La r_{ii} en Conceptos/Vocabulario apareció con las magnitudes más bajas respecto a las demás subescalas. Debido a esta baja inter-relación entre sus ítems, el puntaje obtenido aquí puede ser de cuestionable interpretación si se asume una dimensión latente o ítems con similar contribución. Por otro lado, Percepción Visual fue levemente en esta propiedad bajo pero mejor que la escala anterior. El resto de las escalas mostraron valores r_{ii} alrededor de 0.21, sugiriendo que puedan contener un grado de homogeneidad en las relaciones inter-ítems. Sin embargo, la magnitud de las r_{ii} tendió a mostrar variaciones entre los años y ambos colegios.

La mayoría de las diferencias intra-año en la homogeneidad a través de las subescalas y colegios favorecieron al colegio M.I., ya que la homogeneidad obtenida en el colegio M.I. superaron algunos puntos sobre las correlaciones del colegio S.M.; sin embargo, estas fueron apenas perceptibles debido a su pequeña magnitud y falta de significancia estadística de sus diferencias. Estas diferencias en la homogeneidad pueden sugerir que el desempeño en uno de los colegios puede estar influenciado por factores irrelevantes al constructo, y razonablemente pueden considerados aleatorios.

El ordenamiento de mayor a menor homogeneidad en los ítems fue el siguiente: Letras y Palabras, Conceptos Cuantitativos, Habilidades Fonológicas, Percepción Visual y Conceptos/Vocabulario.

Tabla 28
Homogeneidad (correlación inter-ítems) de los ítems en las subescalas del BDPG, a través de los años y los colegios muestreados.

		2004	2005	2006	2007
Hab. Letras y palabras	MI[1]	0.23	----	0.28	0.23
(18 ítems)	SM[2]	0.28	----	0.24	0.19
	Total	0.28	0.28	0.25	0.21
Hab. Fonológicas	MI	0.17	----	0.24	0.21
(18 ítems)	SM	0.21	----	0.14	0.17
	Total	0.20	0.22	0.19	0.18
Percepción Visual	MI	0.13	----	0.22	0.26
(21 ítems)	SM	0.15	----	0.16	0.11
	Total	0.14	0.24	0.19	0.18
Conceptos Cuantitativos	MI	0.20	----	0.28	0.18
(23 ítems)	SM	0.22	----	0.21	0.14
	Total	0.22	0.25	0.24	0.16
Conceptos/Vocabulario	MI	0.17	----	0.16	0.06
(20 ítems)	SM	0.17	----	0.06	0.15
	Total	0.17	0.15	0.10	0.10

1: Colegio M. I. 2: Colegio SM. En el año 2004, el número de ítems de Conceptos Cuantitativos es 21.

VALIDEZ DE LA ESTRUCTURA INTERNA

Modelamiento conceptual previo

Se probaron varios modelos que podrían explicar las relaciones subyacentes entre los indicadores externos (ítems) y los constructos (variables latentes o factores) con los que teóricamente se asocian. Las restricciones básicas para este análisis fueron: a) cada variable observada se relaciona únicamente con una variable latente, b) no hay covariaciones de error entre las variables observadas, y c) la carga de primera variable observada será fijada con un valor de 1.0, para poder lograr la identificación del modelo.

Los modelos de medición analizados no incluyeron a los ítems de la subescala Conceptos/Vocabulario, pues las evidencias previas mostraron que los ítems parecen no funcionar de manera integrada, mostrando homogeneidad variable y el que más baja correlación inter-ítem promedio contiene. Esta limitación fue consistente en los grupos muestreados (años 2004, 2005, 2006, 2007). Como se observará en los siguientes análisis, los parámetros psicométricos de esta subescala generalmente fueron bajos y señalarían una escala pobremente construida. Por lo tanto, se decidió que esta escala fuera removida del presente análisis para evitar introducir varianza irrelevante al modelamiento.

Las restricciones respecto a la covariación entre las variables latentes (factores) se representarán en los diferentes modelos sometidos al análisis. Estos modelos, representados en la Figura 10 y Figura 11 fueron los siguientes:

Modelo unidimensional. Las respuestas a las tareas fueron modeladas por una estructura parsimoniosa y simple, en que una sola dimensión sería suficiente para explicar la variabilidad de las respuestas y de las relaciones entre las tareas. Se asume errores no correlacionados entre los ítems. Esta única dimensión latente supone un sol factor de primer orden que explica la variabilidad del desempeño en los conglomerados, desde el cual la interpretación del desempeño de los sujetos puede simplificarse en un único puntaje, y que un niño podría mostrar variaciones de rendimiento influenciadas por una sola variable latente.

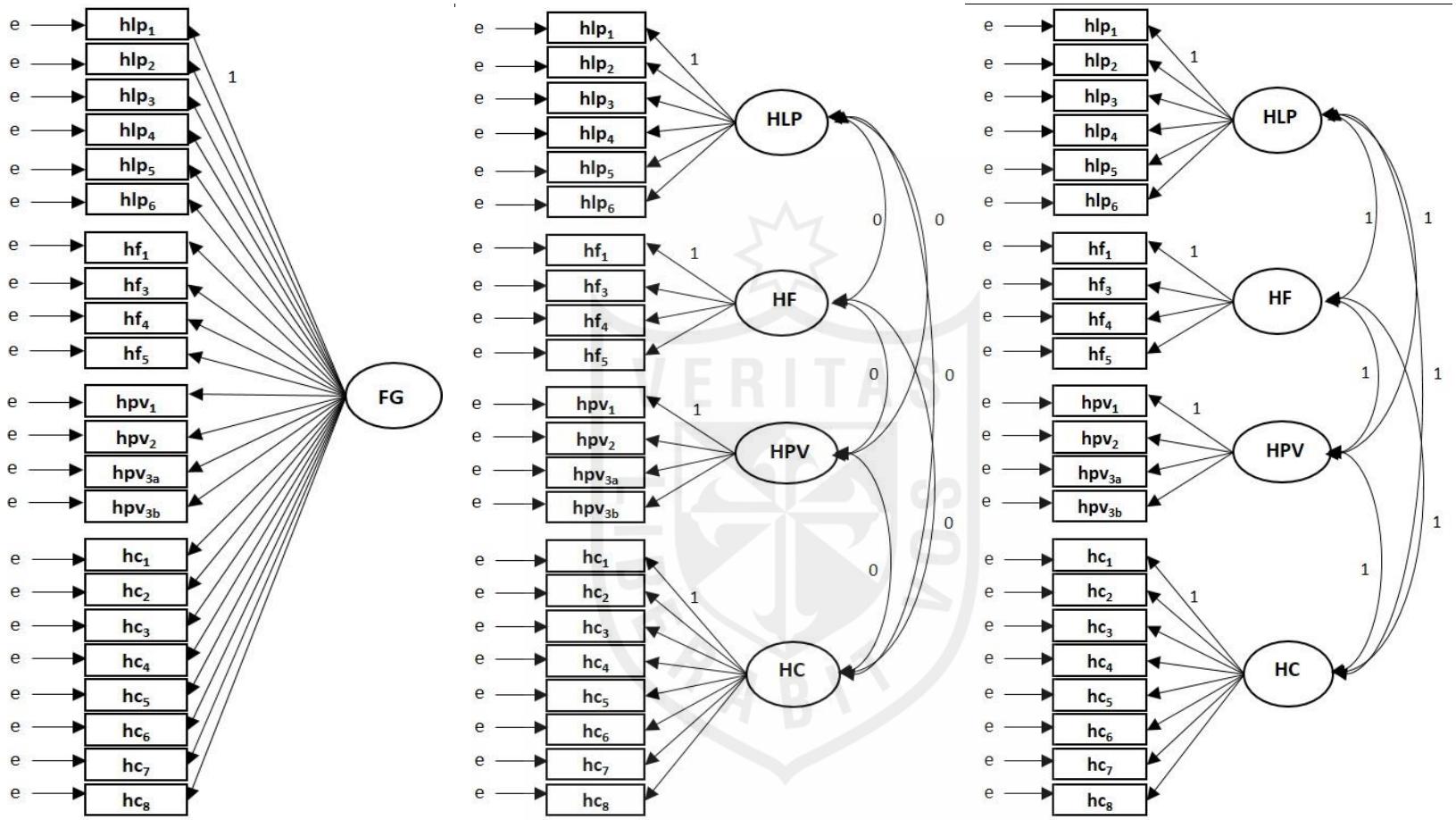


Figura 10
Modelos de medición para el BDPG: Unidimensional, ortogonal y completamente oblicuo

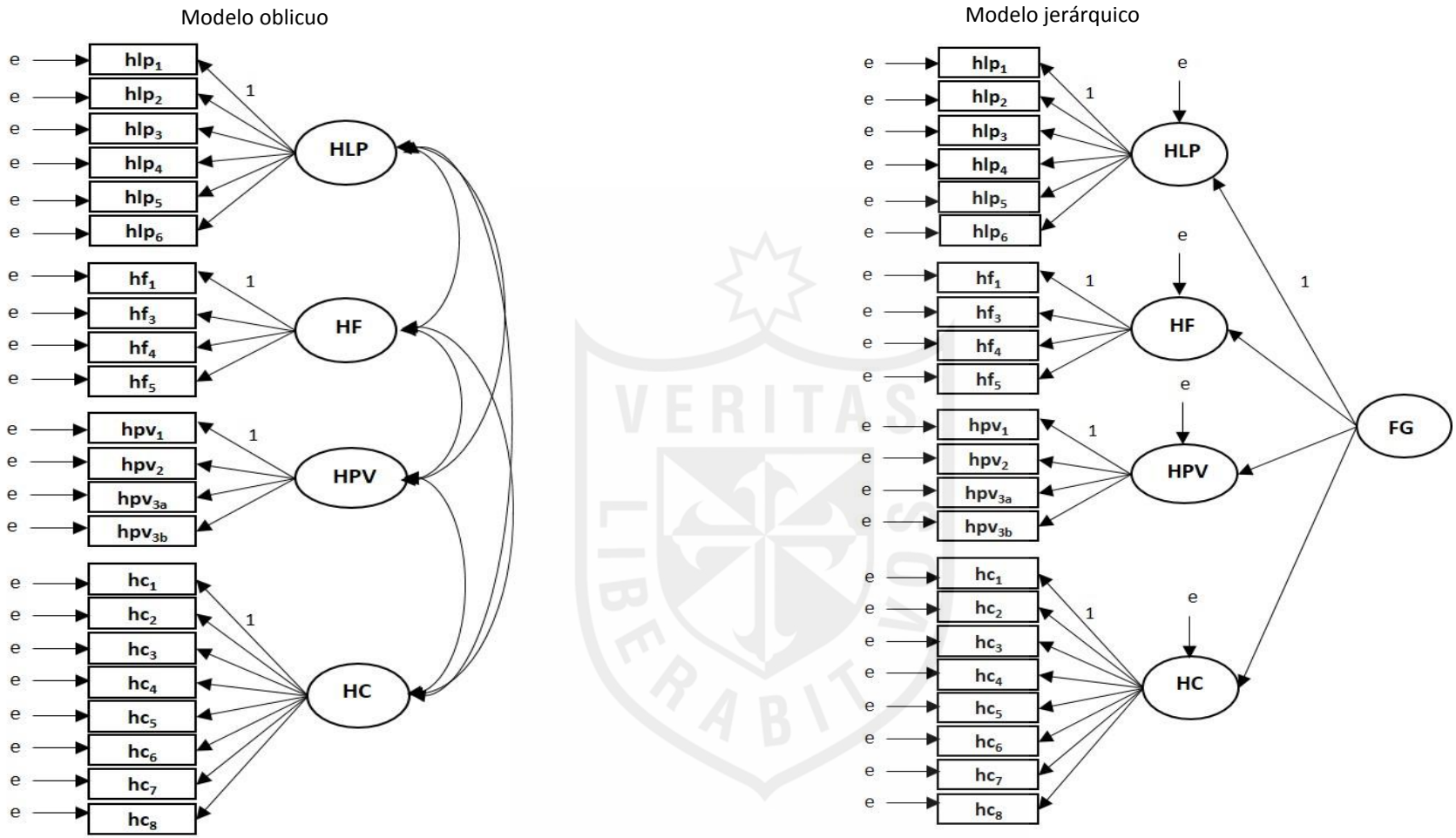


Figura 11
Modelos de medición para el BDPG: Modelo oblicuo y jerárquico

Modelo ortogonal. Aquí, las tareas son modeladas bajo el supuesto de una estructura de partes independientes entre sí, y que las relaciones entre ellas son, por lo tanto, cero. Este modelo supone ausencia de varianza compartida entre los factores, y que la interpretación de un puntaje es completamente independiente del resto. Una estructura como esta supone la ausencia de un puntaje único, y que el desempeño del niño de un grupo de tareas no está asociado su desempeño en otras tareas. Operativamente, los únicos parámetros fijos serán las covarianzas entre los factores, que serán establecidas con cero, mientras que el resto serán estimados libremente. Este modelo es aún parsimonioso, pero relativamente más complejo que el modelo unidimensional.

Modelo completamente oblicuo. En este modelo, las relaciones entre las tareas son completas, indicando redundancia, varianza compartida completa y pobre diferenciación de los contenidos evaluados. En este modelo, la estructura relacional de los factores indica que las tareas no aportan información única sobre el desempeño del niño; por lo tanto, la interpretación de los puntajes no diferenciara contenidos útiles. Operacionalmente, las covarianzas entre los factores serán fijadas con 1.0, mientras que el resto de los parámetros serán estimados libremente. La interpretación del desempeño de un niño dentro de un modelo así expresado podría necesitar solamente un grupo de tareas.

Modelo de factores oblicuos. Este modelo permitirá que los factores muestren algún monto de varianza común, suponiendo que existe un monto de varianza específica suficiente que permita la diferenciación de habilidades, pero también algún monto de varianza compartida. Esta varianza compartida es razonable considerando que las habilidades cognitivas muestreadas en el BDPG no son completamente independientes, y que se influencian en algún monto.

Modelo jerárquico. El último modelo es más complejo, y contendrá varios factores de primer orden y un solo factor de segundo orden. Esta representación jerárquica del modelo de medición del BDPG indica que la comunalidad de los grupos de tareas está explicada por un factor de alto orden, y que expresa una capacidad única y general que subyace a las respuestas de los niños en las tareas del BDPG. Operacionalmente, se creará una variable endógena (factor de segundo orden) que influencia sobre varias variables exógenas (factores de primer orden), todas libremente estimadas. Este modelo supone en primer lugar que las respuestas a las tareas están explicadas por factores diferentes y por factor general, un constructo superior. En segundo lugar, que las correlaciones entre los factores son lo suficientemente altas (pero no muy altas) como para modelar sus covariaciones con la existencia de un factor de segundo orden.

Análisis preliminar

La distribución de la muestra para el análisis de la estructura interna proviene dos colegios y de los años 2005, 2006 y 2007 (Tabla 29). Fueron 461 estudiantes pertenecientes a la muestra original del estudio. Se puede reconocer la proporcionalidad de la muestra en cada año muestreado y entre colegios fue similar. No hay participantes en el colegio SM del año 2005 porque no fueron muestreados.

Tabla 29
Distribución de la muestra de las instituciones educativas MI y SM para el análisis de la estructura interna

	MI		SM		Total	
	N	%	N	%	N	%
2005	107	38.1	-	-	107	23.2
2006	94	33.5	93	51.7	187	40.6
2007	80	28.5	87	48.3	167	36.2
Total	281	100.0	180	100.0	461	100.0

MI y SM son los colegios muestreados

En la Tabla 30 se muestra los clusters o conglomerados de ítems que sirvieron como unidades en la aplicación del método CFA. La agrupación de hizo fundamentalmente tomando en cuenta el contenido de los ítems; esto representaba una estructura construida lógicamente en lugar de usar las correlaciones u otros criterios estadísticos (dificultad o dispersión) entre los ítems. Sin embargo, las características psicométricas de los conglomerados fueron evaluadas. El número de ítems en los conglomerados de Habilidades Fonológicas y de Habilidades de Percepción Visual fueron más homogéneas comparado con Letras y Palabras y Cuantitativa. En todas las agrupaciones, el mínimo número de ítems fue 2, y el máximo número varió entre 3 y 6.

Tabla 30
Definición de agrupaciones de ítems (clusters) y sus ítems componentes

Agrupaciones	Nro de ítems	Items componentes
Habilidades Fonológicas		
HF1	3	p20 + p21 + p22
HF2	3	p23 + p24 + p25
HF3	3	p26 + p27 + p28
HF4	3	p29 + p30 + p31
HF5	4	p32 + p33 + p34+ p35
Habilidades Letras y Palabras		
HLP1	2	p8 + p9
HLP2	4	p10 + p11 + p12 + p13
HLP3	4	p14 + p15 + p16 + p17
HLP4	2	p18 + p19
HLP5	3	p48 + p49 + p50
HLP6	3	p51 + p52 + p53
Habilidad de Percepción Visual		
HPV1	5	p37 to p41
HPV2	6	p42 to p47
HPV3a	5	p54 to p58
HPV3b	5	p59 to p63
Habilidad Cuantitativa		
HC1	2	p64 to p65
HC2	4	p66 to p69
HC3	2	p70 to p71
HC4	2	p72 to p73
HC5	2	p74 to p75
HC6	4	p76 to p79
HC7	2	p80 to p81
HC8	5	p82 to p86

En la

Tabla 31 se muestran los estadísticos descriptivos para los conglomerados. El valor estandarizado de la asimetría y curtosis de los ítems indica una amplia variación en el grado de distorsión de las distribuciones en los conglomerados. La escala con distribución menos asimétrica fue Habilidades Fonológicas, pero en términos absolutos, fue severamente asimétrica (Z > ±2.54). Por lo menos el 50% de conglomerados mostraron severa asimetría (Z > ±3.0). Las escalas tuvieron grupos de tareas con asimetrías positivas y negativas. Esta característica se relacionó con el efecto conjunto de la dificultad de los ítems dentro de ellas. Por otro lado, la curtosis tendió a ser superior a ±3.0, indicando severo alejamiento de una distribución mesocúrtica de las variables. La curtosis multivariada (17.162), y su valor estandarizado (5.43) indican magnitudes claramente alejadas de un valor aceptable de normalidad distribucional multivariada.

Dado la magnitud y la frecuencia con que ocurrieron los excesos en la asimetría y curtosis (asimetría > 2, y curtosis > 7, West et al., 1995), si se aplican métodos de ajuste basados en la teoría normal se estima un serio impacto originado por las características distribucionales de los conglomerados (Curran, et al., 1996; Satorra & Bentler, 1994; West, Finch & Curran, 1995). Esta situación justifica el uso de un método robusto para obtener estimaciones menos sesgadas, desarrollado por Satorra y Bentler (1994).

Tabla 31
Estadísticos descriptivos de las agrupaciones (clusters) de ítems

	M	DE	Min	Max	Asimetría		Curtosis	
					Valor	Z^a	Valor	Z^b
Hab. Fonológicas								
hf1 (3)	1.48	1.068	0	3	0.086	0.750	-1.234	-5.408
hf2 (3)	1.30	1.001	0	3	0.332	2.914	-0.938	-4.112
hf3 (3)	1.89	1.141	0	3	-0.494	-4.328	-1.224	-5.366
hf4 (3)	1.23	1.044	0	3	0.431	3.778	-0.983	-4.308
hf5 (4)	2.36	1.451	0	4	-0.420	-3.682	-1.200	-5.260
Hab. Letras y Palabras								
hlp1 (2)	1.56	.625	0	2	-1.100	-9.643	0.120	0.527
hlp2 (4)	1.89	1.327	0	4	0.093	0.818	-1.149	-5.035
hlp3 (4)	2.08	1.283	0	4	0.124	1.090	-1.064	-4.664
hlp4 (2)	.72	.856	0	2	0.569	4.989	-1.394	-6.110
hlp5 (3)	1.71	1.234	0	3	-0.287	-2.512	-1.531	-6.710
hlp6 (3)	1.68	1.047	0	3	-0.263	-2.306	-1.115	-4.886
Hab. Percepción Visual								
hpv1 (5)	3.16	1.554	0	5	-0.550	-4.824	-0.780	-3.419
hpv2 (6)	2.87	2.008	0	6	0.076	0.662	-1.289	-5.647
hpv3a (5)	3.81	1.522	0	5	-1.168	-10.236	0.363	1.590
hpv3b (5)	3.16	1.538	0	5	-0.597	-5.230	-.595	-2.606
Hab. Cuantitativas								
hc1 (2)	1.20	.789	0	2	-0.373	-3.265	-1.300	-5.697
hc2 (4)	2.66	1.114	0	4	-0.500	-4.383	-0.529	-2.320
hc3 (2)	1.25	.785	0	2	-0.460	-4.030	-1.235	-5.412
hc4 (2)	1.64	.645	0	2	-1.592	-13.952	1.181	5.178
hc5 (2)	1.66	.666	0	2	-1.696	-14.863	1.358	5.954
hc6 (4)	3.00	1.234	0	4	-1.073	-9.401	0.101	0.443
hc7 (2)	1.27	.845	0	2	-0.536	-4.699	-1.387	-6.078
hc8 (5)	2.65	1.487	0	5	-0.071	-0.623	-0.880	-3.855

.[a]: si $Z > \pm 1.96$, $p < 0.05$. .[b]: si $Z > \pm 2.54$, $p < 0.01$

Se aplicaron también correlaciones policóricas entre los conglomerados en lugar de correlaciones Pearson, para atenuar el impacto de la normalidad. Las correlaciones policóricas entre los clusters (Tabla 32) muestran patrones reconocibles sobre la relación entre ellos y las subescalas. El rango de correlación fue desde una valor trivial (0.098) entre conglomerados no asociados, y hasta grande (0.65) entre los conglomerados teóricamente asociados. Tomando en cuenta las sugerencias de Cohen (2001), las correlaciones inter-conglomerados se distribuyen de la siguiente manera en su magnitud:

triviales (menos de 01) = 1 (0.39%), bajas (entre 0.10 y menos de 0.30) = 112 (44.26%), moderadas (entre 0.30 y menos de 0.50) = 133 (52.53%) y altas (desde 0.50) = 7 (2.76%).

El resumen descriptivo y comparativo de los conglomerados en la Tabla 33 revela que la convergencia (correlaciones intra-conglomerado) y divergencia (correlaciones entre conglomerados-) de los conglomerados muestra ser satisfactorio, pues las correlaciones inter-ítem en sus escalas hipotetizadas son comparativamente mayores frente a las escalas no hipotetizadas (Tabla 33). Estos resultados dan soporte para aceptar la validez interna de los conglomerados creados.

También la Tabla 33 muestra la evaluación del ajuste factorial de cada subescala a un modelo unidimensional, que fue exitosa pues se obtuvieron altos coeficientes, llegando incluso a magnitudes inusuales (CFI = 1.00, RMSEA = 0.00). Debido al pequeño número de clusters que sirvieron como indicadores de sus variables latentes, este grado de ajuste no es sorprendente. El método de modelamiento fueron los mismos que se describieron para evaluar el modelo de medición conjunto, que se presenta en la siguiente sección.

Tabla 32

Correlaciones inter-clusters y correlación inter-ítem promedio (en la diagonal)

	hf1	hf2	hf3	hf4	hf5	hlp1	hlp2	hlp3	hlp4	hlp5	hlp6	hpv1	hpv2	hpv3a	hpv3b	hc1	hc2	hc3	hc4	hc5	hc6	hc7	hc8
Hab. Fonológicas																							
hf1	**.281**																						
hf2	**.411**	**.188**																					
hf3	**.110**	**.231**	**434**																				
hf4	**.311**	**.456**	**.335**	**.266**																			
hf5	**.288**	**.274**	**.175**	**.445**	**.395**																		
Hab. Letras y palabras																							
hlp1	.124	.214	.297	.231	.159	**.227**																	
hlp2	.285	.314	.322	.409	.289	**.319**	**.275**																
hlp3	.306	.351	.334	.379	.274	**.328**	**.461**	**.266**															
hlp4	.224	.304	.355	.351	.174	**.313**	**.429**	**.480**	**.595**														
hlp5	.223	.204	.382	.294	.203	**.257**	**.311**	**.391**	**.432**	**.538**													
hlp6	.193	.213	.430	.306	.196	**.327**	**.466**	**.479**	**.511**	**.473**	**.307**												
Hab. Percepción Visual																							
hpv1	.124	.218	.361	.235	.214	.236	.262	.290	.322	.319	.341	**.294**											
hpv2	.318	.435	.292	.385	.303	.320	.392	.523	.554	.290	.396	**.314**	**.344**										
hpv3a	.243	.221	.220	.228	.242	.221	.310	.307	.199	.276	.291	**.306**	**.276**	**.399**									
hpv3b	.209	.224	.288	.222	.258	.258	.219	.347	.286	.298	.319	**.381**	**.360**	**.650**	**.302**								
Hab. Cuantitativas																							
hc1	.261	.297	.300	.343	.235	.278	.355	.351	.306	.301	.334	.352	.330	.304	.328	**1**							
hc2	.273	.282	.282	.300	.242	.282	.344	.345	.309	.326	.338	.332	.323	.205	.287	**.330**	**1**						
hc3	.214	.213	.268	.197	.155	.098	.130	.248	.216	.273	.228	.313	.265	.193	.203	**.180**	**.361**	**1**					
hc4	.214	.210	.333	.266	.202	.331	.280	.337	.276	.318	.336	.297	.291	.275	.347	**.308**	**.338**	**.284**	**1**				
hc5	.226	.139	.276	.219	.200	.293	.278	.295	.259	.292	.331	.290	.272	.295	.358	**.277**	**.310**	**.282**	**.434**	**1**			
hc6	.290	.295	.396	.318	.303	.445	.452	.427	.384	.411	.435	.412	.409	.399	.450	**.439**	**.433**	**.318**	**.537**	**.537**	**1**		
hc7	.281	.352	.346	.399	.237	.405	.476	.449	.401	.393	.404	.372	.445	.380	.410	**.461**	**.435**	**.295**	**.474**	**.403**	**.646**	**1**	
hc8	.443	.346	.251	.327	.302	.203	.331	.407	.317	.326	.353	.325	.378	.282	.270	**.326**	**.400**	**.283**	**.344**	**.286**	**.435**	**.414**	**1**

hf_i: ítem de Habilidades Fonológicas; hlp_i: ítem de Letras y Palabras; hpv_i: ítem de Percepción Visual; hf_i: ítem de Habilidades Cuantitativas.

Tabla 33
Resumen de correlaciones intra-conglomerado y entre-conglomerados

	Correlaciones intra-conglomerado (propia escala)			Correlaciones entre-conglomerado (resto de conglomerados)			CFA		
	M	Min.	Máx.	M	Min.	Máx.	SB-χ^2 (gl)	CFI	RMSEA (IC90%)
Hab. Fonológicas	0.304	0.110	0.456	0.273	0.098	0.443	5.677 (4)	0.99	0.03 (0.0, 0.08)
Hab. Letras y Palabras	0.398	0.257	0.511	0.310	0.098	0.554	9.019 (9)	1.00	0.00 (0.00, 0.05)
Hab. Percepción Visual	0.381	0.276	0.650	0.305	0.124	0.554	0.151 (1)	1.00	0.00 (0.00, 0.09)
Hab. Cuantitativas	0.378	0.180	0.646	0.309	0.098	0.476	32.115* (20)	0.99	0.03 (0.00, 0.05)

M: correlación promedio; Min.: mínima correlación; Máx.; máxima correlación. CFA: análisis factorial confirmatorio. *$p < 0.05$

Análisis del modelamiento CFA

En la Tabla 34, se informa de los residuales para cada modelo ajustado. Los residuales de mayor tamaño se encontraron en el modelo ortogonal y completamente oblicuo, indicando que no se reprodujo bien la relación entre los ítems. El resto de los modelos (unidimensional, oblicuo y jerárquico) fueron los que menores residuales mostraron, y por lo tanto la interpretación de las relaciones entre los ítems puede ser explicados más apropiadamente por la teoría. El promedio de los residuales estandarizados y absolutos son muy similares para estos tres modelos.

Tabla 34
Promedio de los residuales absolutos y estandarizados

Modelos	Valores Absolutos	Valores estandarizados	Rango
Unidimensional	0.05	0.04	-0.2, 0.4
Orthogonal	0.29	0.21	-0.2, >0.5
Completamente oblicuo	0.78	0.58	<-0.5, 0.2
Oblicuo	0.06	0.04	-0.2, 0.3
Jerárquico	0.04	0.04	-0.1, 0.3

En la Tabla 35 se muestran los resultados de la evaluación del ajuste de los modelos. Antes de pasar a describir los resultados debe anotarse que la diferencia entre la estimación del método asintótico normal y el robusto fue de considerable magnitud. Esto señala que el impacto de la desviación de la normalidad multivariada produciría infraestimación de los índices de ajuste y parámetros con error estándar inflados.

Tabla 35
Resultados del AFC y medidas de ajuste para los modelos de medición

Modelo (gl)	χ^2	AIC	SRMR	RMSEA[I.C. 90%]	CFI
A priori					
Unidimensional (230)					
Asintótica	1285.664**	825.664	0.062	0.100 [0.09, 0.11]	0.94
Robusto SB	673.556**	213.556		0.065 [0.05, 0.07]	0.97
Orthogonal (230)					
Asintótica	2224.187**	1764.187	0.319	0.10 [0.10, 0.11]	0.89
Robusto SB	1275.017**	815.018		0.09 [0.09, 0.11]	0.93
Complt. Oblicuo (253)					
Asintótica	1035.462**	575.462	0.89	0.08 [0.08, 0.09]	0.78
Robusto SB	1010.959**	550.959		0.08 [0.08, 0.09]	0.78
Oblicuo (224)					
Asintótica	994.10**	546.101	0.065	0.08 [0.08, 0.09]	0.95
Robusto SB	506.482**	58.48		0.05 [0.04, 0.06]	0.98
Jerarquico (226)					
Asintótica	979.924**	527.924	0.064	0.08 [0.08, 0.09]	0.85
Robusto SB	476.759**	24.759		0.04 [0.04, 0.05]	0.94
A posteriori[a]					
Oblicuo (223)					
Asintótica	877.633**	431.633	0.05	0.08 [0.07, 0.08]	0.96
Robusto SB	443.758**	-2.241		0.04 [0.04, 0.05]	0.98
Jerarquico (225)					
Asintótica	846.245**	396.244	0.05	0.07 [0.07, 0.08]	0.88
Robusto SB	408.034**	-41.965		0.04 [0.03, 0.04]	0.95

[a]Reespecificacion liberando el término de error entre HPB3a y HPB3b. **p < 0.01

El modelo unidimensional, ortogonal y completamente oblicuo fueron los que tuvieron peores índices de ajuste, especialmente los dos primeros. En el caso unidimensional, esto señala que un hay varianza única que supera debe ser modelada, pues existen variaciones en las respuestas de acuerdo al tipo de tarea presentada. Con respecto al modelo ortogonal, el desajuste esencial proviene de no tomar en cuenta la covariación natural entre las variables latentes. Finalmente, el fracaso del modelo redundante (completamente oblicuo) proviene del exceso de covariación representada.

El modelo oblicuo mostró un excepcional buen ajuste comparado con los modelos anteriores, ubicándose en la zona de compromiso satisfactorio en cada uno de los índices de ajuste. La correlación interfactorial, estima en este modelo aparece en la Tabla 36. Las correlaciones superan en valor 0.65, y claramente muestran fuerte varianza compartida entre ellas.

El modelo jerárquico produjo un leve mejor ajuste respecto al modelo oblicuo, y ambos producen un nivel de ajuste que puede considerarse satisfactorio. Se revisó el índice de Lagrange para explorar posibles mejoras a los modelos oblicuo y jerárquico, y hallar parámetros que conceptualmente y estadísticamente no fueron considerados. Esta fase del análisis es una evaluación a posteriori de los modelos, y por lo tanto de naturaleza exploratoria se re-especificaron los dos modelos de mejor ajuste y representación teórica apropiada: modelo oblicuo y jerárquico. Para el modelo oblicuo, se halló que se necesitaba liberar la covarianza específica entre dos conglomerados de la escala de Habilidad Perceptual Visual; esta re-especificación produciría diferencias sustanciales en la prueba de bondad de ajuste ($\Delta SB\text{-}\chi^2 = 128.37$, $p < 0.001$).

En el modelo jerárquico, esta re-especificación también producirían cambios estadísticamente significativos ($\Delta SB\text{-}\chi^2 = 139.25$, $p < 0.001$). Por lo tanto, se re-especificó

la covariación entre los errores de ambos clusters. Los resultados aparecen en la Tabla 35, debajo del encabezado A posteriori. Los índices mejoraron, pero levemente. La covariación entre los términos de error liberados fue estadísticamente significativa, (z = 8.97, error estándar = 0.043, r = 0.558, *Cov* = 0.383).

Los parámetros de los ítems aparecen en la Tabla 36. Excepto en la escala de Habilidades Cuantitativas, el rango de las cargas factoriales de los ítems va entre 0.50 y 0.85. En general, puede considerarse que las cargas factoriales de los ítems son adecuadas y el poder discriminativo es adecuado para la interpretación del constructo. La magnitud de la confiabilidad de los ítems generalmente supera el valor 0.30, lo que puede tomarse como un valor adecuado de varianza explicada de cada parcela. La confiabilidad del constructo, ρ (Fornell & Larcker, 1981; Werts Rock, & Linn, 1978), basado en las cargas factoriales, se halla en niveles adecuados ($\geq$ 0.70, Fornell & Larcker, 1981). La excepción ocurrió con HPV, cuyo valor estuvo levemente bajo, pero luego de aplicar la corrección por errores correlacionados (Raykov, 2001), predeciblemente disminuyó aún más (ρ = 0.59).

Tabla 36
Parámetros estandarizados de los ítems

	HF	HLP	HPV	HC	G	R^2
hf1	0.524					0.274
hf2	0.631					0.398
hf3	0.567					0.321
hf4	0.729					0.531
hf5	0.554					0.307
hlp1		0.597				0.356
hlp2		0.669				0.447
hlp3		0.737				0.544
hlp4		0.801				0.642
hlp5		0.659				0.435
hlp6		0.740				0.548
hpv1			0.562			0.316
hpv2			0.675			0.455
hpv3a			0.544			0.296
hpv3b			0.576			0.332
hc1				0.627		0.393
hc2				0.605		0.366
hc3				0.483		0.234
hc4				0.744		0.553
hc5				0.713		0.508
hc6				0.859		0.738
hc7				0.853		0.727
hc8				0.607		0.369
Factores						
HF					0.844	0.712
HLP					0.917	0.841
HPV					0.992	0.985
HC					0.923	0.852
ρ	0.74	0.85	0.68 (0.59)	0.88	0.95	
Correlaciones[a]						
HF	1					
HLP	0.772	1				
HPV	0.698	0.756	1			
HC	0.777	0.847	0.816	1		

[a]Correlaciones estimadas con el modelo oblícuo. ρ = confiabilidad del constructo; entre paréntesis, esimtacion corregida por correlaciones entre los términos de error de las parcelas hpv3a y hpv3b.G: factor de segundo orden.

VALIDEZ DE CONSTRUCTO: RELACIONES DIVERGENTES Y CONVERGENTES

Los estadísticos descriptivos básicos para mostrar estas evidencias de validez se hicieron en varios momentos y grupos. Las relaciones divergentes con las pruebas de desarrollo (TEPSI) y aptitudes escolares (PDP) se presentan primero, y cuyos estadísticos descriptivos aparecen en la Tabla 37. Las relaciones divergentes con las medidas de habilidad intelectual (DAP: IQ y TONI-2) se presentarán en segundo lugar, con los estadísticos descriptivos mostrados en la Tabla 40. Finalmente, las correlaciones con medidas de habilidad visomotora (VMI y Bender - QSS) se presentarán en último lugar; los estadísticos descriptivos para este grupo aparecen en la Tabla 42.

Relación divergente una prueba de aptitudes preescolares (PDP)

Aplicando el enfoque de prueba de hipótesis tradicional para la correlaciones y el de Braden (1986), la Tabla 38 muestra los valores de las correlaciones basadas en H_o: $\rho = 0$, y las correlaciones obtenidas con la confiabilidad conjunta entre las subescalas correlacionadas (H_1: $\rho = 1$). Las correlaciones estadísticamente significativas bajo la columna H_1 señalarán que las subescalas no son lo suficientemente altas como para sugerir equivalencia y similaridad en el atributo latente medido; este tipo de resultados daría un componente interpretativo orientado hacia la divergencia entre las subescalas BDPG y las de PDP; contrariamente, si no se rechaza H1, entonces ello sugiere equivalencia y similaridad dado que la correlación obtenida es similar a la máxima correlación posible, al menos en un sentido cuantitativo.

Tabla 37
Estadísticos descriptivos basados en los puntajes directos para PDP, BDPG y TEPSI[a]

	Puntajes directos						KR21
	Mín.	Máx.	M	D.E.	Asimetría[1]	Curtosis[2]	
PDP							
Conceptos Verbales	3	16	11.16	3.88	-.532	-1.005	0.85
Conceptos Cuantitativos	1	14	9.74	4.56	-.871	-.868	0.92
Memoria Auditiva	2	6	4.39	1.30	-.500	-.687	0.0
Coordinación Motora	6	12	9.94	2.09	-.976	-.308	0.82
Percepción Visual	19	49	33.68	8.78	.170	-.799	0.71
PDP total (Sin Mem. Aud.)	43	94	68.90	15.51	-.058	-.935	0.75[3]
BDPG							
Letras-Palabras	2	18	9.52	4.60	.519	-.495	0.86
Hab-Fon	1	17	9.87	4.28	.255	-.285	0.84
Hab. Percepción Visual	5	20	13.16	4.45	.023	-1.187	0.82
Vocabulatio/Conceptos	7	20	15.35	3.60	-.830	-.016	0.78
Hab. Cuantitativas	4	22	15.35	5.65	-.746	-.972	0.88
Total	30	93	63.25	17.88	-.006	-.737	0.84[3]
TEPSI							
Coordinación	4	16	13.94	2.863	-1.981	4.350	0.83
Lenguaje	8	23	16.35	4.013	-.119	-.623	0.71
Motricidad	6	12	10.81	1.424	-1.560	3.075	0.52
Total	22	51	41.10	6.764	-.776	.793	0.63[3]

[a]: Participantes del PRONOEI (año 2007), n = 31. [1]: Error estándar = 0.42 [2]: error estándar = 0.82. [3]: Confiabilidad basado en el coeficiente alfa estratificado.

Respecto al puntaje de Letras y Palabras, este ha convergido con los puntajes de Aptitud Perceptual Visual, pero también con Conceptos Verbales y Conceptos Cuantitativos en un nivel alto (r_{prom} = 0.59), y sugiere que la discriminación y reconocimiento de letras y palabras comparten alrededor de 34.8% de varianza común con las habilidades perceptuales visuales, el vocabulario comprensivo global y específico (cuantitativo). Con estas dos últimas aptitudes del PDP, estas correlaciones parecen no satisfacer la hipótesis de divergencia. Sin embargo, probando contra la máxima

correlación posible (columna H_1), se rechazó tal hipótesis ($p < 0.01$), lo que demuestra que no son medidas equivalentes. Por otro lado, su correlación con Coordinación Motora da cuenta de la esperada divergencia máxima (cero) con esta habilidad medida.

Sobre las Habilidades Fonológicas, este ha mostrado también elevada varianza común ($r_{prom} = 0.55$) con Conceptos Verbales y Aptitud Perceptual Visual, sugiriendo la inter-influencia de estas aptitudes. Sin embargo, estas correlaciones no indican una identidad cuantitativa con estas aptitudes, pues fueron no se acercaron a su máxima correlación posible con ellas ($p < 0.01$ ó $p < 0.05$). De otro lado, se demuestra su divergencia de constructo con Conceptos Cuantitativos y Coordinación Visomotora, ya que las magnitudes de las correlaciones son cercanas a cero y provenientes del error de muestreo ($p > 0.05$).

La subescala de Hab. de Percepción Visual ha mostrado un claro patrón de correlaciones convergentes respecto a Aptitud Perceptual Visual del PDP, ya que con esta medida tuvo la más alta correlación; esto converge esperablemente ya que ambos se orientan hacia el mismo constructo pero con tareas diferentes. Por otro lado, las asociaciones divergentes estuvieron entre niveles provenientes del azar ($p > 0.05$) con Conceptos Cuantitativos y Coordinación Motora ($r_{prom} = 0.18$). También, su delimitación de contenido (cuantitativo) se verificó al observar que su correlación con Aptitud Perceptual Visual del PDP es similar a la máxima correlación posible con ella ($p > 0.05$).

Tabla 38

Correlaciones convergentes y divergentes entre el BDPG y Prueba de Diagnóstico Preescolar (PDP) [a]

BDPG	Hab. Letras y Palabras		Hab. Fonológicas		Hab. Percepción Visual		Hab. Vocabulario/ conceptos		Hab. Cuantitativas		BDPG Total	
PDP	H_o[1]	H_1[2]	H_o	H_1	H_o	H_1	H_o	H_1	H_o	H_1	H_o	H_1
Conceptos Verbales	0.669**	0.85**	0.522**	0.84**	0.487**	0.83**	0.470**	0.81**	0.658**	0.86**	0.727**	0.84
Conceptos Cuantitativos	0.522**	0.88**	0.249	0.87**	0.300	0.86**	0.432*	0.84**	0.657**	0.89	0.577**	0.87**
Coordinación Motora	0.004	0.83**	0.193	0.82**	-0.110	0.82**	-0.043	0.79**	-0.187	0.84**	-0.053	0.82**
Aptitud Perceptual Visual	0.596**	0.78*	0.580**	0.77*	0.675**	0.76	0.453*	0.74**	0.546**	0.79**	0.724**	0.77
PDP Total	0.689**	0.80**	0.583**	0.79*	0.61**	0.78**	0.52**	0.76**	0.67**	0.81*	0.79**	0.79

[a]Participantes del PRONOEI (año 2007), n = 31. [1]Prueba de hipótesis contra $\rho = 0$. [2]Prueba de hipótesis contra $\rho = 1$ (enfoque Braden, 1986).
**$p < 0.01$; *$p < 0.05$

Considerando la subescala Vocabulario/Conceptos, su convergencia con Conceptos Verbales ha sido menos de lo que se podría esperar para medidas orientadas hacia el mismo fin; este nivel de asociación es similar a medidas divergentes como Percepción Visual. Con Coordinación Motora, en las que se muestra su divergencia de constructo. Su diferencia con la máxima correlación da cuenta también de la no equivalencia con las medidas de aptitudes escolares.

Finalmente, Habilidades Cuantitativas converge apropiadamente con Conceptos Cuantitativos, pero el tamaño de las correlaciones con medidas divergentes del PDP es similar (r_{prom} = 0.57), excepto con Coordinación Motora (esperablemente bajo). Comparado con la máxima correlación posible entre ambas, la aceptación de la hipótesis alternativa sugiere que ambas subescalas son equivalentes.

Relación divergente con prueba de desarrollo psicomotor (TEPSI)

La covariación con la medida de desarrollo psicomotor ha sido satisfactorias (Tabla 39) de acuerdo a la hipótesis de divergencia, ya que van por debajo del nivel moderado ($r_{máx}$ = 0.44). Ya que no se asume que el BDPG mida el desarrollo psicomotor general del niño que ingresa a primer grado, su puntaje total correlaciona moderadamente bajo con las subpruebas de Coordinación y Lenguaje. Aunque la varianza común entre ambas es estadísticamente significativa en este (como en las otros análisis correlacionales presentados), más bien ocurren entre las medidas con las que aún se espera que compartan contenidos dado los constructos. Por ejemplo, Letras y Palabras se asocia con Lenguaje y Coordinación de TEPSI ($p < 0.05$); estas subpruebas del TEPSI evalúan aspectos de abstracción, discriminación colores, etc.; y Coordinación contiene también elementos de abstracción ya que tiene al dibujo de la figura humana entre sus ítems.

Por otro lado, Percepción Visual comparte varianza común con Coordinación ya que posee tareas manipulativas que requieren habilidades perceptuales visuales y control motor, ambos involucrados en la función de integración visomotora (Beery, 2000; Brannigan & Decker, 2003). La aplicación del enfoque de Braden (1986) obtuvo resultados satisfactorios contra todos los puntajes del TEPSI, ya que se rechazó la hipótesis de igualdad con la máxima correlación posible. Esto indica que aunque las relaciones lineales entre el BDPG y el TEPSI contienen alta variabilidad sistemática entre ellas, están aún lejanas del criterio que sugiere que ambas medidas para considerarlas como medidas convergentes.

Relación divergente con pruebas de inteligencia (DAP: IQ y TONI-2)

Los resultados descriptivos se presentan en la Tabla 40, y se muestra la tendencia hacia la normalidad de los puntajes, específicamente de las medidas de inteligencia, y especialmente del DAP: IQ. El resto de los puntajes (del BDPG) muestran también un aceptable acercamiento a la distribución normal teórica, excepto Vocabulario/Conceptos, que presentó severa asimetría negativa en los datos del año 2006; por lo tanto, los resultados correlacionales con esta subescala deben interpretarse con precaución. Los análisis presentados en la Tabla 41 indican covariaciones lineales entre los puntajes del BDPG y las medidas de inteligencia más allá de valor nulo ($r = 0$). Entre el puntaje total del BDPG y ambas medidas de habilidad intelectual, se hallaron correlaciones de magnitud similar ($\chi^2 < 3.6$, $p > 0.05$) entre los años. Así mismo, las correlaciones de las subpruebas con las medidas intelectuales también fueron similares, y estuvieron alrededor de 0.39. El nivel de asociación lineal hallada indica un moderado nivel de varianza compartida.

Tabla 39
Correlaciones divergentes entre el BDPG y el TEPSI [a]

BDPG / TEPSI	Hab. Letras y Palabras		Hab. Fonológicas		Hab. Percepción Visual		Hab. Vocabulario/ conceptos		Hab. Cuantitativas		BDPG Total	
	H_o[1]	H_1[2]	H_o	H_1	H_o	H_1	H_o	H_1	H_o	H_1	H_o	H_1
Coordinación	0.364*	0.84**	0.284	0.83**	0.419*	0.82**	0.306	0.80**	0.440*	0.85**	0.465**	0.83**
Lenguaje	0.425*	0.78**	0.328	0.77**	0.352	0.76**	0.232	0.74**	0.203	0.79**	0.387*	0.77**
Motricidad	0.194	0.66**	0.184	0.66**	0.150	0.65**	0.200	0.63**	0.192	0.67**	0.223	0.66**
TEPSI Total	0.448*	0.74**	0.356*	0.73**	0.424*	0.73**	0.308	0.71**	0.347	0.75**	0.475**	0.73**

[a]Participantes del PRONOEI (año 2007), n = 31. [1]Prueba de hipótesis contra $\rho = 0$. [2]Prueba de hipótesis contra $\rho = 1$ (enfoque Braden, 1986).
**$p < 0.01$; *$p < 0.05$

Todas las correlaciones fueron estadísticamente significativas ($p < 0.01$), y de acuerdo a la prueba de comparación entre correlaciones independientes, estas difieren por error de muestreo. Con el DAP: IQ, las correlaciones han tendido a una mayor variación entre sí, pero estas diferencias no han sido lo suficientemente grandes como para ser detectadas por la prueba χ^2. La varianza compartida entre las subpruebas del BDPG y las medidas de habilidad intelectual es ligeramente mayor con el DAP:IQ; en contraste, con el TONI-2 fue algo menor. Con el DAP, la varianza compartida en el año 2006 varía entre 7% y 20%, y en el año 2007 entre 7% y 26%; mientras que con los puntajes del TONI-2, varía entre 9% y 14%. Al aplicar una prueba chi cuadrado para evaluar la diferencia entre varias correlaciones independientes (entre los años), no se detectaron diferencias estadísticamente significativas, y se obtuvo la media ponderada de las correlaciones en los años muestreados ($\bar{X}_{pond}$), que mue. Se puede observar que todas las correlaciones promedio obtenidas tienen magnitudes similares (Tabla 41).

Tabla 40
Estadísticos descriptivos basados en los puntajes directos para BDPG y las medidas de habilidad intelectual (DAP: IQ y TONI-2)

	2006[a]				2007[b]			
	M	DE	As	Cu	M	DE	As	Cu
Letras-Palabras	8.3	4.8	0.4	-0.9	9.9	4.3	0.04	-0.92
Hab-Fonológicas	8.4	3.9	0	-0.5	9.6	3.9	0.03	-0.76
Hab. Percepción Visual	13.3	4.6	-0.5	-0.2	14.1	4.4	-0.68	0.09
Hab. Cuantitativas	15	5.6	-0.7	-0.3	15.6	4.5	-0.39	0.93
Vocabulario/Conceptos	14.5	3.4	-1.1	1.8	14.7	3.3	-0.66	0.21
BDPG	72.8	19.7	-0.5	0.2	76.9	17.7	-0.2	-0.4
DAP: IQ[b]	15.6	5.2	0.4	0.7	16.6	4.9	0.001	-0.6
TONI-2[c]	-	-	-	-	9.6	3.5	0.12	-0.47

[a]Tamaño muestral: n = 180. [b]Tamaño muestral para el DAP:IQ en el 2007, n = 149. [c]Tamaño muestral para el TONI-2 en el 2007, n = 87.

Tabla 41
Correlaciones lineales entre las subescalas y puntaje total del BDPG y las medidas de habilidad intelectual (DAP:IQ y TONI-2)

	TONI 2007 (n = 87)	DAP: IQ 2006 (n = 180)	DAP: IQ 2007 (n = 149)	$\bar{X}_{pond}$	$\chi^2_{(2)}$
Letras-Palabras	0.42**	0.44**	0.38**	0.41	0.42
Hab-Fonológicas	0.34**	0.51**	0.37**	0.42	3.55
Hab. Percepción Visual	0.28**	0.41**	0.31**	0.35	1.67
Hab. Cuantitativas	0.34**	0.40**	0.36**	0.37	0.33
Vocabulatio/Conceptos	0.45**	0.28**	0.31**	0.33	2.29
BDPG	0.47**	0.53**	0.44**	0.49	1.16
$\bar{X}$	0.38	0.43	0.36		

** $p < 0{,}01$. $\bar{X}_{pond}$: promedio inter-correlaciones ponderado por el tamaño muestral de cada grupo. $\chi^2_{(2)}$: prueba de diferencia de correlaciones independientes.

Relación convergente con pruebas de habilidades visomotoras (Bender QSS y VMI-4)

Se correlacionaron las pruebas del BDPG con medidas de habilidad visomotora, buscando obtener evidencias de las relaciones divergentes entre ellas. Los estadísticos básicos se encuentran en la Tabla 42. Los estadísticos presentados en esta tabla muestran similitud cuantitativa, y podrían ser asumidos como valores constantes entre los años. Luego, el análisis de las correlaciones presentado en la

Tabla 43 muestra que todas las correlaciones fueron estadísticamente significativas, sugiriendo que las mismas pueden ser explicadas por variaciones sistemáticas entre ambos constructos en la población desde cual provienen las muestras. Los datos sumarios indican que las correlaciones con el Bender-QSS tienden a ser más elevadas que con los puntajes del VMI.

Las correlaciones lineales entre las medidas de habilidad visomotora y la BDPG demuestran una varianza compartida que varía entre 7% y 16% para el VMI, y entre 17% y 34% para en Bender-QSS; esta variación entre los coeficientes ha sido un poco mayor con el VMI (rango = 0.21) que con el Bender–QSS (rango = 0.17). Como consecuencia de la magnitud de estas correlaciones, el promedio de las correlaciones fue mayor con el Bender-QSS. Al examinar las diferencias entre las correlaciones aplicando una prueba *Z* para correlaciones independientes, mayormente estas no fueron estadísticamente significativas, excepto las correlaciones en la subescala Letras-Palabras y Habilidad de Percepción Visual. Sin embargo, luego de aplicar la corrección Bonferroni a los valores *p* (ver

Tabla 43), la correlación en Habilidad de Percepción Visual puede ser consideradas como efecto del error de muestreo debido al efecto de las múltiples pruebas de significancia hechas. Por lo tanto, la validez diferencial entre el QSS y el VMI ocurrió en la subescala Letras-Palabras.

Tabla 42
Estadísticos descriptivos basados en los puntajes directos para BDPG y medidas visomotoras (VMI-4 y Bender – QSS)

	2006 (n = 56)				2007 (n = 87)			
	M	DE	As	Cu	M	DE	As	Cu
Letras-Palabras	7.8	5.0	0.6	-0.6	9.5	4.2	0.2	-0.7
Hab. Fonológicas	8.3	4.5	0.1	-0.9	9.4	3.9	0.1	-0.7
Hab. Percepción Visual	13.2	5.3	-0.5	-0.3	14.0	3.7	-0.2	-0.6
Hab. Cuantitativas	14.3	6.3	-0.7	-0.4	15.0	4.3	-0.6	0.3
Vocabulario/Conceptos	14.3	4.5	-1.2	1.3	14.3	3.8	-0.6	-0.2
BDPG	71.6	23.9	-0.7	0.5	74.1	17.0	0.1	-0.5
VMI-4	-	-	-	-	10.7	2.1	-0.8	1.3
Bender – QSS	18.7	3.1	-0.6	-0.06	-	-	-	-

Tabla 43

Correlaciones lineales entre las subescalas y puntaje total del BDPG y medidas visomotoras (VMI-4 y Bender – QSS)

	VMI-4 (n = 87)	Bender - QSS (n = 56)	Z [b]
Letras-Palabras	0.27*	0.59**	-2.28[c]
Hab-Fonológicas	0.35**	0.53**	-1.281
Hab. Percepción Visual	0.28**	0.55**	-1.885[d]
Hab. Cuantitativas	0.48**	0.53**	-0.383
Vocabulario/Conceptos	0.40**	0.42**	-0.137
BDPG	0.47**	0.64**	-1.414
$\overline{X}$ [a]	0.356	0.524	
DE [a]	0.0873	0.0631	

*: $p < 0.05$. **: $p < 0.01$. [a]La media y la desviación estándar calculadas no incluyeron el BDPG. [b]p ajustado por corrección Bonferroni: alfa/Nro comparaciones = 0.05/6 = 0.008. [c]$p = 0.001$. [d]$p = 0.03$.

VALIDEZ CONCURRENTE

Los estadísticos descriptivos básicos para mostrar estas evidencias de validez se hicieron en varios momentos y grupos. Los estadísticos descriptivos de BDPG y la Prueba Gestáltica Anton Brenner (PGAB) se presentan en la Tabla 44. La prueba t de Student aplicada a la diferencias entre los puntajes de los dos grupos muestreados (2006 y 2007), indica la similaridad en la respuesta promedio, excepto para el puntaje total del BDPG. El rendimiento en este puntaje total es alto en el 2007, así como su dispersión. Esto está influenciado por el rendimiento en Letras y Palabras, pues su asimetría distribucional en el año 2007 es cerca de cero, pero en el año 2006 es fuerte. Otra discrepancia notoria ocurre en el rendimiento con la PGAB, que es elevado en grupo del año 2007; la dispersión del puntaje es, en cambio casi la mitad menos que en el año 2006. Esta discrepancia puede relacionarse con el efecto de la variabilidad de los estadísticos producido por el pequeño tamaño muestral en el 2006.

Tabla 44
Estadísticos descriptivos de los puntajes directos para BDPG y PGAB

	2006 (n = 26)				2007 (n = 80)				T[a](*d*)
	M	D.E.	As[1]	Cu[2]	M	D.E.	As[3]	Cu[4]	
BDPG									
Letras-Palabras	8.73	5.31	-1.91	3.76	9.92	4.33	0.05	-0.92	1.15 (0.23)
Hab. Fonológicas	9.00	4.56	0.35	-0.94	9.67	3.95	0.03	-0.77	0.72 (0.16)
Hab. Percepción Visual	13.08	5.33	-0.04	-0.69	14.19	4.45	-0.69	0.09	1.05 (0.24)
Hab. Cuantitativas	15.15	5.59	-0.76	0.52	15.65	4.57	-0.40	0.93	0.45 (0.10)
Vocabulario/Conceptos	13.85	4.54	-0.82	0.58	14.73	3.31	-0.67	0.22	1.07 (0.24)
Total	59.81	21.91	-1.05	1.90	64.16	16.47	-0.18	-0.27	1.07 (0.24)
PGAB	68.92	11.08	-0.68	0.84	71.04	5.99	-1.37	3.01	1.24 (0.28)

[1]Error estándar = 0.45. [2]Error estándar = 0.88. [3]Error estándar: 0.27. [4]Error estándar: 0.53. [a]Grado de libertad para el *t* de Student = 104. *d*: diferencia estandarizada. BDPG: Batería de Despistaje para Primer Grado. PGAB; Prueba Gestáltica de Antonn Brenner.

Las correlaciones ajustadas por el método Olkin-Pratt (1958) fueron levemente superiores las correlaciones no corregidas. Por otro lado, se hallaron correlaciones moderadas y fuertes entre la PDPG y PGAB, pero se reconoce un patrón en la magnitud de las mismas: las correlaciones tienden a ser más altas entre el PGAB y las pruebas de percepción visual y habilidades cuantitativas; este patrón se repite en ambas muestras. Cualitativamente, estas correlaciones son mayores a 0.50 y pueden considerarse grandes. En contraste, las correlaciones entre PGAB y Letras-Palabras y habilidades fonológicas son menores, específicamente de nivel moderado.

Para probar la replicabilidad de las correlaciones en ambas muestras, la comparación estadística realizada por la prueba *Z* para correlaciones independientes, arrojó que ninguna de las comparaciones fue estadísticamente significativa. Aunque la correlación en la prueba de Vocabulario/Conceptos fue notoriamente diferente entre los grupos, no alcanzó significancia estadística debido al bajo poder estadístico. Estos resultados indican que las correlaciones pueden considerarse similares.

Tabla 45
Correlaciones entre BDPG y Prueba Gestáltico de Antonn Brenner (PGAB), en dos grupos

	Correlaciones con PGAB		*Z*[a]
	2006 (n = 26)	2007 (n = 80)	
Letras-Palabras	0.446* (0.454)	0.440**	0.03
Hab-Fonológicas	0.494* (0.502)	0.441**	0.28
Hab. Percepción Visual	0.667** (0.675)	0.517**	0.98
Hab. Cuantitativas	0.631** (0.639)	0.523**	0.68
Vocabulario/Conceptos	0.534** (0.542)	0.306**	1.17
BDPG	0.710** (0.718)	0.582**	0.93

*: $p < 0.05$. **: $p < 0.01$. [a]Prueba *z* para la diferencia de correlaciones independientes. Entre paréntesis, correlaciones ajustadas por sesgo de pequeña muestra (método Olkin & Pratt, 1958). BDPG: Batería de Despistaje para Primer Grado. PGAB; Prueba Gestáltica de Antonn Brenner.

VALIDEZ POR DIFERENCIACIÓN DE GRUPOS

Los estadísticos descriptivos y la consistencia interna para los puntajes en la muestra total se exhiben en la Tabla 46. Las estimaciones de consistencia interna presentadas en esta tabla se basaron en varios coeficientes que consideraron el modelo de medición subyacente a los puntajes (Cronbach, 1951; Gilmer-Feldt, 1983; Horst, 1951); de este modo la estimación de la confiabilidad se hizo más precisa. Los resultados de estas estimaciones sugieren que el monto de error de medición es aceptable en esta muestra, y que puede tener solo un impacto leve en las pruebas estadísticas (Charter, 1997) que se aplicarán en este objetivo de análisis.

Las pruebas que evaluaron el funcionamiento perceptual (Prueba de Percepción Visual No Motor y Prueba Gestáltica de Bender – Brannigan y Brunner) tendieron a ser algo bajas, pero el resto de las pruebas obtuvieron coeficientes de ≥ 0.80. En la Tabla 46, también se muestra la estimación de la media y desviación estándar para la muestra total (n = 40), mientras que DE_{total} corresponde a la estimación en la muestra completa de niños del cual proviene los grupos extremos (ver ANEXO C).

Aplicando las técnicas inferenciales basados en el método de Feldt (1961) y Peters (1941; Thorndike, 1989), en la Tabla 47 se observa que, exceptuando la prueba DAP: IQ y Conceptos/Vocabulario, las diferencias fueron estadísticamente significativas.

Tabla 46
Información descriptiva y consistencia interna de los puntajes

	Consistencia interna	Alto rendimiento (n = 21)		Bajo Rendimiento (n = 19)		Total		
		M	DE	M	DE	M	DE	DE_{total}[4]
Prueba de Percepción Visual	0.71[1]	18.33	3.979	15.68	4.922	17.07	4.593	4.724
CI (DAP:IQ)	0.80[2]	17.05	5.352	15.42	5.242	16.28	5.296	5.387
Prueba Gestáltica Brenner	0.84[3]	69.10	6.602	62.37	12.361	65.90	10.220	10.332
Prueba Gestáltica de Bender (Brannigan-Brunner)	0.71[3]	20.19	1.778	19.37	2.793	19.80	2.323	2.352
BDPG								
Letras y Palabras	0.92[1] (0.88[3])	13.67	3.610	8.16	4.549	11.05	4.899	4.923
Conciencia Fonológica	0.85[1] (.80[3])	11.29	4.014	8.26	3.034	9.85	3.853	3.888
Percepción. Visual	0.88[1] (0.84[3])	15.62	4.387	12.89	4.748	14.33	4.709	4.761
Comprensión/Vocabulario	0.83[1] (0.79[3])	13.90	3.727	13.05	4.390	13.50	4.026	4.077
Cuantitativo	0.86[1] (0.83[3])	17.10	3.300	11.95	5.060	14.65	4.918	4.949
Total		71.75	14.421	54.32	16.97	67.28	18.688	17.936

1: coeficiente KR-20 con modificación Horst; 2: coeficiente Gilmer-Feldt; 3: coeficiente alfa de Cronbach o KR-20; 4: valores estimados obtenido del método de Garret (1971) (ver ANEXO C)

Las correlaciones de magnitud del efecto halladas fueron moderadas o altas. La corrección de las correlaciones por restricción del rango (Preacher, 2007) hicieron disminuir algunas correlaciones estimadas por el método de Feldt (1961), pero se elevaron consistentemente con el método de Peters (1944). Ya que con este método ocurrió un incremento con mejor sentido lógico, la interpretación final se hizo con estas estimaciones. Todas las subpruebas (excepto Conceptos/Vocabulario) mostraron efectos altos (o moderado en Percepción Visual) sobre la diferenciación de grupos. Algo llamativo que la fuerza de la correlación en la pruebas de percepción visual fueron similar. También, la prueba PGB también estuvo alrededor de la magnitud de la correlaciones en la BDPG.

Por otro lado, la magnitud de las diferencias estandarizadas (*g Hedges*) tendió a ser de mayor intensidad en los puntajes de la BDPG, comparado con las otras pruebas, algo que repite el patrón ocurrido en las correlaciones. De manera esperada, la excepción también ocurrió en la subprueba Conceptos/Vocabulario, ya que los grupos obtuvieron puntajes similares. Claramente, las subpruebas en que las diferencias fueron más fuertes son Letras y Palabras, Conciencia Fonológica y Habilidades Cuantitativas, que representan conjuntos que directamente se asocian al rendimiento en la lectura y las matemáticas. El puntaje total también logra diferenciar claramente a los dos grupos, lo que es razonable pues con tiene el efecto acumulativo de la capacidad discriminativa de las subescalas.

Tabla 47
Resultados de la comparación entre grupos

	Fedlt (1961)		Peters (1944)		*.g Hedges*
	t.	*r.*	$r_{no\text{-}cor}$	r_{cor}	
Prueba de Percepción Visual no Motor	1.776	0.078	0.207	0.474*	1.06
CI (DAP:IQ)	0.931	0.039	0.11	0.374	0.79
Prueba Gestáltica Antonn Brenner	2.055[a]	.084	0.237	0.503*	1.14
Prueba de Bender (Brannigan-Brunner)	1.06	0.045	0.127	0.385*	0.82
BDPG					
Letras y Palabras	4.06**	0.14	0.407**	0.643**	1.65
Conc. Fonológica	2.577**	0.10	0.28**	0.544**	1.27
Perc. Visual	1.807	0.074	0.20	0.435**	0.95
Comp./Vocab.	0.631	0.027	0.075	0.314	0.65
Cuantit.	3.648**	0.134	0.379***	0.622**	1.56
Total	3.349**	0.125	0.354***	0.546**	1.28

[a.]: Debajo del nivel p con ajuste Bonferroni, p/nro de comparaciones = 0.05/4 = 0.012
**:Debajo del nivel p con ajuste Bonferroni, p/nro de comparaciones = 0.05/6 = 0.008.
$r_{no\text{-}cor}$: correlación no corregida; $r_{no\text{-}cor}$: correlación corregida

CONFIABILIDAD (CONSISTENCIA INTERNA)

La modificación de Horst elevó sistemáticamente los coeficientes KR-20, pero los cambios cuantitativos no parecen ser de gran magnitud, ya que estas elevaciones representaron generalmente un 5% respecto al KR-20 no atenuado. Estos cambios significaron que todos los coeficientes corregidos estuvieran más cerca del nivel 0.90, y que por efectos del redondeo pueden alcanzarlo. Estas magnitudes se considerarán más apropiadas ya que afinan mejor la precisión de la estimación de la consistencia interna calculada en los puntajes de la presente muestra; por lo tanto, en el siguiente reporte sobre las comparaciones *intra-año* (Tabla 48) y *entre-año* (Tabla 49), se usarán los coeficientes modificados. En la Tabla 48 y Tabla 49 se muestran las estimaciones de confiabilidad corregidas y no corregidas; además de las pruebas de hipótesis para probar la igualdad de los coeficientes, que solo se efectuó en los coeficientes corregidos. El nivel será ajustado por el método Bonferroni.

Comparaciones intra-año de la consistencia interna

Para el reporte principal de este análisis, se tomarán las estimaciones de confiabilidad KR-20 corregidas por Host. Los resultados mostraron que los coeficientes de confiabilidad, corregidos y no corregidos, fueron moderadamente variables entre los colegios (Tabla 48); por otro lado, se usó la corrección Bonferroni para controlar la inflación del error Tipo I sobre las comparaciones entre coeficientes. En el año 2004, los coeficientes fueron numéricamente cercanos, y en ninguno demostró ser estadísticamente diferente ($p < 0.05$); esta afirmación se cumplió en todas las subescalas en ese año. En el año 2006, las subescalas Habilidades Fonológicas y Conceptos/Vocabulario fueron desigualmente consistentes pues los coeficientes fueron mayores para MI y las diferencias estadísticamente significativas. Por lo tanto, entre ambos colegios y para estas subescalas, la consistencia interna no se consideró equivalente. En el año 2007, únicamente las

subescalas Percepción Visual y nuevamente Conceptos/Vocabulario mostraron diferencias más allá del error de muestreo, pero la dirección de las diferencias fue opuesto: los puntajes de los niños del colegio MI fueron más consistentes que los del colegio SM en Percepción Visual, mientras los niños del SM produjeron puntajes más consistentes en Conceptos/Verbales que los niños del colegio MI.

Exceptuando el año 2004, generalmente las estimaciones de confiabilidad fueron superiores para el colegio MI, pues en el año 2006 las confiabilidades en el colegio MI fueron entre 1% y 43% más altos con respecto a los coeficientes provenientes del colegio SM; y las mayores diferencias provinieron en la subescala Habilidades Fonológicas y Conceptos/Vocabulario, pues en ambas los puntajes en el colegio MI fueron estadísticamente diferentes a los del colegio SM en estas subescalas. Y en el año 2007, las diferencias entre los coeficientes fueron entre 2% y 72% más elevados en el colegio MI respecto al colegio SM; las mayores discrepancias provinieron de las subescalas Hab. Fonológicas, Percepción Visual y Conceptos/Vocabulario, en las cuales se detectó significancia estadística como se mencionó en un párrafo anterior. En contraste con lo anterior, en que los puntajes derivados del colegio MI fueron relativamente más confiables, en el año 2004 los coeficientes en el colegio SM fueron apenas superiores respecto al colegio MI, pues esta superioridad fue alrededor del 4%.

Se puede reconocer algunos grupos de escalas que fueron más consistentes que otras, por ejemplo, las subescalas Hab. Letras y Palabras fueron los más consistentes a través de las comparaciones intra-año; y el segundo grupo de subescalas consistentes fue la integrada por Hab. Fonológicas y Percepción Visual, que fueron menores pero aun altamente confiables. Finalmente, Conceptos/Vocabulario fue la subescala que comparativamente fue de más baja confiabilidad, que incluso llegó a 0.70 (año 2006).

Cualitativamente, las subescalas estuvieron generalmente por sobre el límite de los mínimos niveles de consistencia interna recomendados para decisiones aplicadas, por lo que pueden ser evaluados como satisfactorios para propósitos del uso del BDPG. En conclusión, exceptuando una escala (Conceptos/Vocabulario), la consistencia interna en las estimaciones intra-año fueron generalmente satisfactorias, y las comparaciones entre los colegios sugirieron que las estimaciones de confiabilidad pueden considerarse moderadamente similares. El patrón de diferencias generalmente tendió a favorecer a un colegio sobre el otro, aunque las diferencias fueron generalmente pequeñas, Estas confiabilidades pueden, por lo tanto, ser estimadas para la muestra total dentro de cada año y llegar a una estimación general en cada año, que puede ser afectada por poca variabilidad. El nivel cualitativo de las estimaciones de confiabilidad fue satisfactorio y tendió a tener pocas variaciones, lo que es esperable se acuerdo a las pocas variaciones cuantitativas halladas entre los coeficientes de confiabilidad

Comparaciones entre-años de la consistencia interna

Como se observa en la Tabla 49, las estimaciones de los coeficientes con y sin la modificación de Horst repiten el patrón hallado en el análisis anterior: los coeficientes corregidos fueron mayores, y aunque esta superioridad no fue cuantitativamente grande, sin embargo representa una mejor estimación. Los siguientes resultados tomarán como referencia los coeficientes corregidos.

Tabla 48

Comparación de la consistencia interna intra-años: KR20 y KR20 con ajuste de Horst

		2004		2006		2007	
Subescala	Colegio	.α	α_{Horst}	.α	α_{Horst}	.α	α_{Horst}
Hab. Letras y palabras (18 ítems)							
	MI	0.85	0.87	0.88	0.90	0.84	0.87
	SM	0.87	0.90	0.86	0.89	0.81	0.85
	$.\chi^2(1)$	0.42	1.43	0.48	0.18	0.54	0.37
	.p	0.51	0.23	0.48	0.65	0.46	0.54
Hab. Fonológicas (17 ítems)							
	MI	0.78	0.81	0.85	0.87	0.82	0.86
	SM	0.82	0.86	0.74	0.76	0.72	0.77
	$.\chi^2(1)$	0.83	1.92	6.13	7.59	3.51	4.42
	.p	0.36	0.16	0.01	0.005	0.06	0.03
Percepción Visual (21 ítems)							
	MI	0.75	0.77	0.85	0.88	0.87	0.90
	SM	0.78	0.80	0.80	0.82	0.72	0.75
	$.\chi^2(1)$	0.34	0.41	3.42	3.42	10.59	14.91
	.p	0.55	0.52	0.06	0.06	0.001	< 0.001
Conceptos Cuantitativos[a] (24 ítems)							
	MI	0.84	0.86	0.89	0.92	0.82	0.86
	SM	0.86	0.88	0.86	0.88	0.79	0.82
	$.\chi^2(1)$	0.37	0.50	1.22	3.45	0.44	1.18
	.p	0.53	0.47	0.26	0.06	0.50	0.27
Conceptos/Vocabulario[b] (20 ítems)							
	MI	0.76	0.79	0.79	0.82	0.56	0.59
	SM	0.75	0.79	0.54	0.57	0.78	0.81
	$.\chi^2(1)$	0.03	0.00	12.51	15.33	8.82	10.82
	.p	0.85	1.0	< 0.001	< 0.001	0.003	0.001

[a]basado en 21 ítems en el año 2004. [b]basado en 14 ítems en el año 2004

Exceptuando esta última subescala, usando la modificación de Horst, el resto estuvo sobre 0.83, que identifica el nivel consistencia como de buen nivel de control del error de medición y que podrían sugerir el uso para decisiones individuales. Los pequeños cambios en la confiabilidad se observó en la prueba de hipótesis para diferencias en la

confiabilidad entre los años, pero únicamente para Hab. Letras y Palabras, y Vocabulario/Conceptos. En el resto de los puntajes, las estimaciones de consistencia interna entre-años fueron estadísticamente significativas (para todas las comparaciones, $p < 0.01$). A través de los años, Hab. Letras y Palabras, y Conceptos Cuantitativos fueron los puntajes levemente más confiables, seguidos por Percepción Visual y Hab. Fonológicas; en la Tabla 49 también se puede comprobar que Conceptos/Vocabulario fue sistemáticamente baja respecto a los demás puntajes.

El patrón de diferencias entre las pruebas a través de los años indica que en el año 2005 la consistencia interna de los puntajes tendió a ser mayor, y en el año 2007 esta tendió a ser menor. La prueba chi cuadrado para comparar confiabilidades independientes arrojó que, mientras que Hab. Letras y Palabras, y Conceptos/Vocabulario muestran ser internamente consistentes a través de los años, en los demás puntajes se han detectado diferencias estadísticamente significativas. A continuación, las comparaciones pareadas entre cada año sobre cada escala, se hicieron ajustando por el método Bonferroni. Específicamente las diferencias significativas entre estos años provienen de los puntajes de Hab. Fonológicas en el año 2005 frente a los estimados en el año 2007 (χ^2 (1) = 11.59, $p < 0.01$) y 2006 (χ^2 (1) = 10.01, $p < 0.01$). En Percepción Visual, el coeficiente calculado en el año 2005 fue superior al estimado en el año 2004 ($\chi^2 = 15.37$, $p < 0.01$). Finalmente, la consistencia interna de Conceptos Cuantitativos fue diferente únicamente entre las estimaciones del año 2005 y 2007, $\chi^2(1) = 11.20$, $p < 0.01$.

Tabla 49

Comparación de la consistencia interna entre-años: KR20 y KR20 con ajuste de Horst, para la muestra total en cada año

	2004		2005		2006		2007		$\bar{X}_{\alpha}$		$\chi^2(3)$
	$.\alpha$	α_{Horst}	$.\alpha$	α_{Horst}	$.\alpha$	α_{Horst}	$.\alpha$	α_{Horst}	$.\alpha$	α_{Horst}	
Hab. Letras y palabras	0.87	0.89	0.87	0.90	0.87	0.89	0.83	0.86	0.86	0.89	4.04 ($p = 0.25$)
Hab. Fonológicas	0.82	0.84	0.82	0.89	0.80	0.82	0.77	0.81	0.80	0.84	12.60 ($p < 0.001$)
Percepción Visual	0.77	0.78	0.86	0.89	0.83	0.84	0.82	0.84	0.82	0.84	16.21 ($p < 0.01$)
Conceptos Cuantitativos	0.85	0.87	0.88	0.91	0.88	0.90	0.81	0.83	0.86	0.88	15.97 ($p > 0.05$)
Conceptos/ Vocabulario	0.76	0.78	0.78	0.80	0.70	0.72	0.71	0.73	0.74	0.76	5.34 ($p = 0.14$)

En conclusión, la consistencia interna ha tendido a ser estable entre los años; y aunque se han detectado diferencias entre algunos años, estos parecen representar un evento específico e idiosincrásico de la variación encontrada. Esta inestabilidad en los coeficientes se relacionó a los puntajes obtenidos en un año de recolección de los datos, en que las estimaciones fueron moderadamente más elevadas. Por otro lado, exceptuando la subescala Conceptos/Vocabulario, las estimaciones de consistencia interna generalmente pasaron el límite de 0.80.

CARACTERÍSTICAS DISTRIBUCIONALES DE LOS PUNTAJES

Gradiente y Piso para Letras/Palabras.

Gradiente. En la distribución de puntajes de la subescala Letras/Palabras (Tabla 50), los tres primeros puntajes generalmente se ubican alrededor de -1.5 D.E. debajo de la media, y la gradiente en ese nivel no ha sido adecuada, así como en el máximo puntaje posible. En todos los años ha ocurrido este problema, sin embargo esto no se expandió a todos los niveles de los puntajes: desde el puntaje 4, la gradiente fue modalmente entre 2 y 3 puntos de diferencia entre los puntajes, y frente al mínimo establecido (5 puntos o menos) ello sugiere que la gradiente del ítem fue adecuada para los puntajes situados entre -1.0DE y 1.5DE sobre la media. Entre el puntaje 0 y 1, y luego 17 y 18, han ocurrido los mayores problemas de gradiente, pues la magnitud del cambio fue mayor a lo esperado, llegando a una diferencia de 29 puntos (año 2007). En el nivel bajo de desempeño, este problema sin embargo no ha sido constante, pues una parte de todos los puntajes no tuvo una gradiente adecuada. Se puede concluir que la escala Letras/Palabras, los puntajes debajo de -1 desviación estándar de la distribución de puntajes son parcialmente adecuadas en la medición de niños con bajo rendimiento en esta escala.

Piso. De acuerdo al criterio para un piso apropiado (puntaje estandarizado de 60 o menos), únicamente la distribución del año 2005 y 2007 tuvieron un piso apropiado; en el año 2006, el piso estuvo muy cerca del criterio. En el año 2004 (Tabla 50), el piso estuvo sobre el criterio, y por lo tanto, no fue adecuado. En conjunto, esta evidencia sugiere que el piso de esta subescala es parcialmente estable, y pues alcanzar el límite inferior de 2 DE debajo de la media que puede ser interpretable.

Tabla 50
Gradiente del ítem para la subescala Letras/Palabras en los años muestreados

PD	2004			2005			2006			2007		
	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]
18	132	-	-	133	-	-	133	-	-	131	-	-
17	124	8	No	124	9	No	125	8	No	123	8	No
16	119	5	Sí	120	4	Sí	121	5	Sí	119	4	Sí
15	116	3	Sí	117	3	Sí	117	4	Sí	116	3	Sí
14	113	3	Sí	114	2	Sí	115	2	Sí	113	3	Sí
13	111	2	Sí	113	1	Sí	112	2	Sí	109	3	Sí
12	108	2	Sí	111	2	Sí	110	2	Sí	107	3	Sí
11	106	2	Sí	108	3	Sí	108	2	Sí	103	3	Sí
10	104	2	Sí	106	3	Sí	106	2	Sí	100	3	Sí
9	102	2	Sí	103	3	Sí	104	2	Sí	97	3	Sí
8	100	3	Sí	101	2	Sí	101	3	Sí	95	2	Sí
7	97	3	Sí	99	1	Sí	97	3	Sí	93	2	Sí
6	94	2	Sí	97	2	Sí	95	3	Sí	89	4	Sí
5	91	3	Sí	94	4	Sí	92	3	Sí	84	5	Sí
4	87	4	Sí	89	4	Sí	89	3	Sí	80	5	Sí
3	83	5	Sí	84	6	No	85	4	Sí	75	4	Sí
2	78	5	Sí	76	8	No	79	6	No	69	6	No
1	73	5	Sí	69	7	No	70	9	No	40	29	No
0	64	9	No	61	8	No	58	12	No	40	0	Sí

[a]Diferencia entre el puntaje correspondiente al resultado de la celda y el puntaje siguiente superior.
[b]Resultado de comparar los dos puntajes consecutivos y el criterio usado (Sí cumple o No cumple con el criterio)

Gradiente y Piso para Habilidades Fonológicas.

Gradiente. Como en la escala anterior, la gradiente del ítem para Habilidades Fonológicas estuvo de acuerdo con el criterio entre +- 1 D.E, pero inadecuados para los puntajes entremos en ambas colas de la distribución (Tabla 51). En la región que más podría interesar (la zona del rendimiento debajo del promedio ubicado en el extremo inferior), entre los años 2005 y 2007 hubo menos problemas de gradiente en comparación con el año 2004. En este último grupo, existe una pobre gradiente debajo del puntaje

estandarizado 80, lo que produciría poca precisión descriptiva de los niños en este nivel de puntaje.

Piso. El piso de la distribución en los años 2004 y 2007 fueron apenas apropiados, pues aunque estuvieron debajo del criterio establecido, únicamente existen entre uno y tres puntajes directos para su interpretación nomotética. Los puntajes en los años 2005 y 2007, el criterio para el piso apropiado no se cumple.

Tabla 51
Gradiente del ítem para la subescala Habilidades Fonológicas en los años muestreados

PD	2004			2005			2006			2007		
	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]
17	131	-	-	133	-	-	135	-	-	131	-	-
16	123	8	No	126	7	No	128	6	No	122	10	No
15	118	5	No	121	5	No	123	5	Sí	117	5	Sí
14	114	4	Sí	116	4	Sí	120	4	Sí	114	3	Sí
13	112	2	Sí	113	3	Sí	117	3	Sí	112	3	Sí
12	110	2	Sí	110	3	Sí	113	4	Sí	109	3	Sí
11	106	3	Sí	106	4	Sí	110	4	Sí	105	4	Sí
10	102	4	Sí	102	3	Sí	106	4	Sí	101	4	Sí
9	99	3	Sí	99	3	Sí	102	4	Sí	98	3	Sí
8	96	3	Sí	96	3	Sí	98	4	Sí	95	3	Sí
7	93	3	Sí	93	3	Sí	94	3	Sí	91	4	Sí
6	89	3	Sí	90	3	Sí	92	3	Sí	87	4	Sí
5	86	3	Sí	87	3	Sí	88	3	Sí	83	4	Sí
4	81	5	Sí	83	4	Sí	84	5	Sí	79	4	Sí
3	75	6	No	79	4	Sí	79	5	No	75	4	Sí
2	69	6	No	76	3	Sí	75	3	Sí	69	7	No
1	62	7	No	70	5	No	72	3	Sí	40	29	No
0	40	22	No	61	9	No	65	7	No	40	0	-

[a]Diferencia entre el puntaje correspondiente al resultado de la celda y el puntaje siguiente superior.
[b]Resultado de comparar los dos puntajes consecutivos y el criterio usado (Sí cumple o No cumple con el criterio)

Gradiente y Piso para Habilidad de Percepción Visual.

Gradiente. La gradiente del ítem para Habilidades de Percepción visual fue apropiada entre el rango de +1.5DE y -2DE; y alrededor de la media, la gradiente tendió a presentar 3 unidades de desviación estándar de diferencia entre los puntajes (Tabla 52), específicamente en los años 2004, 2005 y 2006. Por el contrario, la mejor gradiente de esta subescala se halló en el año 2007, pues entre -1DE y -2DE la diferencia constante entre los puntajes estandarizados fue dos unidades de DE. Esta constancia no ocurrió en los otros años, y podría deberse al mejoramiento del proceso de construcción y de administración de las pruebas, así como de las características idiosincrásicas de las muestras. Por otro lado, los problemas de la gradiente del ítem en esta subescala son similares a los otras subescalas, es decir, que los puntajes alrededor de lo que representan un alto y bajo desempeño, muestras gradientes inapropiadas, especialmente en este último nivel de habilidad. En este nivel, el salto de un puntaje a otro llega a una diferencia de más de 20 unidades de DE en los años 2004 y 2007.

Piso. El piso de la distribución en los años 2004, 2006 y 2007podrían considerarse satisfactorios, pues se ubican a 4DE debajo de la media e indican desempeños claramente pobres en estas áreas (Tabla 52). Por el contrario, el piso de este puntaje no fue satisfactorio en el año 2005, pues el puntaje estandarizado correspondiente al puntaje 1 fue mayor a 69. En este año, el desempeño fue mejor que en los otros periodos y por lo tanto el puntaje fue más asimétricamente negativo. Este evento puede representar, sin embargo, un comportamiento eventual de la muestra recolectada, pero indica una inestabilidad en identificar un piso consistentemente apropiado para los puntajes de esta escala.

Tabla 52
Gradiente del ítem para la subescala Habilidad de Percepción Visual en los años muestreados

PD	2004			2005			2006			2007		
	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]
21	135	-	-	139	-	-	133	-	-	133	-	-
20	128	7	No	128	11	No	125	9	No	123	10	No
19	122	6	No	119	9	No	118	6	No	116	7	No
18	118	4	Sí	114	5	No	114	4	Sí	111	5	Sí
17	114	4	Sí	111	3	Sí	110	3	Sí	107	4	Sí
16	110	4	Sí	108	3	Sí	107	3	Sí	104	3	Sí
15	106	4	Sí	105	3	Sí	103	4	Sí	101	3	Sí
14	103	3	Sí	101	4	Sí	100	4	Sí	98	3	Sí
13	100	3	Sí	98	3	Sí	97	3	Sí	95	3	Sí
12	97	3	Sí	96	3	Sí	94	2	Sí	92	3	Sí
11	94	3	Sí	92	3	Sí	92	3	Sí	88	4	Sí
10	91	3	Sí	90	3	Sí	89	3	Sí	85	3	Sí
9	87	4	Sí	88	1	Sí	86	3	Sí	83	2	Sí
8	83	4	Sí	87	2	Sí	83	3	Sí	80	2	Sí
7	81	2	Sí	84	2	Sí	80	3	Sí	79	2	Sí
6	77	4	Sí	83	2	Sí	78	2	Sí	77	2	Sí
5	73	5	Sí	81	1	Sí	76	2	Sí	75	2	Sí
4	70	2	Sí	80	1	Sí	74	2	Sí	72	2	Sí
3	68	3	Sí	77	3	Sí	71	3	Sí	70	2	Sí
2	62	6	No	75	3	Sí	68	3	Sí	64	6	No
1	40	22	No	73	2	Sí	64	4	Sí	40	24	No
0	40	0	Sí	67	6	No	58	6	No	40	0	Sí

[a]Diferencia entre el puntaje correspondiente al resultado de la celda y el puntaje siguiente superior.
[b]Resultado de comparar los dos puntajes consecutivos y el criterio usado (Sí cumple o No cumple con el criterio)

Gradiente y Piso para Habilidades Cuantitativas

Gradiente. En la Tabla 53, la gradiente de la distribución de puntajes de esta escala fue adecuada para una gran parte de ella. En todos los años, la gradiente del ítem estuvo fue menos que el máximo valor sugerido (es decir, una diferencia de 5 unidades estandarizadas). La gradiente fue mejor entre -1DE y -2DE de los puntajes, pues la

diferencia entre los puntajes estandarizados fue alrededor de 2 y 3 unidades DE; incluso, en el año 2005, hubo una gradiente de 1 punto en algunos niveles de puntuación. Por otro lado, la gradiente se hizo mucho menos adecuada en los extremos de la distribución; esto ocurrió en todos los periodos evaluados, pero particularmente en los puntajes mayores de ±2DE de la media.

Piso. El piso de los puntajes de esta escala fue adecuado en todos los años muestreados. En todos los periodos, el piso fue menor que el nivel exigido, y el mínimo puntaje posible pudo ser alcanzado por la muestra.

Tabla 53
Gradiente del ítem para la subescala Habilidades Cuantitativas en los años muestreados

PD	2004			2005			2006			2007		
	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]
24	-	-	-	135	-	-	-	-	-	-	-	-
23	-	-	-	123	13	No	133	-	-	138	-	-
22	138	-	-	115	8	No	125	8	No	125	13	No
21	127	11	No	111	4	Sí	119	6	No	119	6	No
20	120	8	No	107	4	Sí	113	6	No	115	4	Sí
19	115	5	Sí	104	3	Sí	108	5	Sí	110	4	Sí
18	110	4	Sí	102	2	Sí	104	4	Sí	107	3	Sí
17	106	4	Sí	100	2	Sí	100	4	Sí	104	3	Sí
16	102	4	Sí	97	3	Sí	98	2	Sí	100	4	Sí
15	99	3	Sí	95	3	Sí	97	2	Sí	96	4	Sí
14	96	3	Sí	93	2	Sí	95	2	Sí	92	3	Sí
13	94	3	Sí	91	2	Sí	93	2	Sí	89	3	Sí
12	91	3	Sí	89	2	Sí	92	1	Sí	87	2	Sí
11	89	2	Sí	87	2	Sí	90	2	Sí	84	2	Sí
10	88	1	Sí	86	2	Sí	87	2	Sí	82	2	Sí
9	87	1	Sí	84	1	Sí	85	2	Sí	80	3	Sí
8	85	2	Sí	84	1	Sí	84	2	Sí	77	3	Sí
7	82	3	Sí	82	1	Sí	81	2	Sí	76	1	Sí
6	79	3	Sí	79	3	Sí	79	2	Sí	74	2	Sí
5	75	3	Sí	76	3	Sí	77	2	Sí	69	4	Sí
4	73	3	Sí	74	1	Sí	74	3	Sí	66	3	Sí
3	71	2	Sí	72	2	Sí	72	2	Sí	66	0	Sí
2	67	4	Sí	70	2	Sí	70	1	Sí	62	4	Sí
1	58	9	No	67	3	Sí	68	3	Sí	40	22	No
0	40	18	No	61	6	No	62	6	No	40	0	Sí

[a]Diferencia entre el puntaje correspondiente al resultado de la celda y el puntaje siguiente superior.
[b]Resultado de comparar los dos puntajes consecutivos y el criterio usado (Sí cumple o No cumple con el criterio)

Gradiente y Piso para Habilidades de Vocabulario

Gradiente. En los años muestreados, el gradiente no fue adecuado en los puntajes sobre la media (

Tabla 54), específicamente en el puntaje estandarizado de 103 o más en los años 2005, y 2006. También, la gradiente fue inapropiada en ambos extremos de la distribución de los puntajes en el año 2004 y 2007. Se observa que en el año 2005 y 2006, la gradiente fue más variable que en los puntajes de los años 2004 y 2007.

Piso. Este aspecto parece ser apropiado para para este puntaje, pues los niños alcanzaron el mínimo rendimiento posible (puntaje 1) debajo del criterio establecido. Sin embargo, en el año 2007, el puntaje mínimo no es 1 sino 4. La muestra ha mostrado mejor habilidad de vocabulario en este año, y el piso se ha iniciado en cerca de 2DE debajo del promedio. Los resultados de este atributo psicométrico se muestran inestables pues no han sido similares en los años muestreados.

Tabla 54
Gradiente del ítem para la subescala Habilidades de Vocabulario en los años muestreados

PD	2004			2005			2006			2007		
	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]
24	-	-	-	-	-	-	-	-	-	-	-	-
23	-	-	-	-	-	-	-	-	-	-	-	-
22	-	-	-	-	-	-	-	-	-	-	-	-
21	-	-	-	-	-	-	-	-	-	141	-	-
20	-	-	-	135	-	-	135	-	-	131	11	No
19	-	-	-	125	10	No	127	8	No	123	8	No
18	136	-	-	117	8	No	117	10	No	114	8	No
17	128	8	No	109	8	No	110	7	No	108	6	No
16	121	7	No	103	6	No	104	6	No	104	4	Sí
15	115	6	No	100	3	Sí	100	4	Sí	100	4	Sí
14	111	5	Sí	97	3	Sí	96	4	Sí	95	5	Sí
13	107	4	Sí	95	3	Sí	92	5	Sí	91	5	Sí
12	103	4	Sí	91	4	Sí	88	4	Sí	87	3	Sí
11	100	3	Sí	87	5	Sí	85	3	Sí	84	3	Sí
10	97	3	Sí	83	4	Sí	82	3	Sí	80	4	Sí
9	94	3	Sí	80	3	Sí	77	5	No	75	4	Sí
8	91	3	Sí	78	2	Sí	73	4	Sí	73	2	Sí
7	88	3	Sí	76	2	Sí	70	2	Sí	71	2	Sí
6	84	4	Sí	73	3	Sí	70	1	Sí	69	2	Sí
5	78	6	No	71	2	Sí	68	2	Sí	64	5	Sí
4	70	8	No	69	3	Sí	65	2	Sí	40	24	No
3	64	6	No	65	4	Sí	65	0	Sí	40	0	No
2	58	6	No	65	0	Sí	64	2	Sí	40	0	No
1	40	18	No	65	0	Sí	62	2	Sí	40	0	No
0	40	0	Sí	61	4	Sí	58	3	Sí	40	0	No

[a]Diferencia entre el puntaje correspondiente al resultado de la celda y el puntaje siguiente superior.
[b]Resultado de comparar los dos puntajes consecutivos y el criterio usado (Sí cumple o No cumple con el criterio)

Gradiente y Piso para el Puntaje Total

Gradiente. La gradiente del puntaje total puede considerarse excelente (Tabla 55), pues en todos los niveles de puntuación la gradiente cumplió con el criterio previamente definido. La excepción fue, sin embargo, en el puntaje final superior e inferior de cada escala. El problema de la gradiente puede no ser problema en estos niveles de puntuación, pues las hay pocos cambios en los entre puntajes consecutivos debido que son estimados en muestras pequeñas en estas zonas de puntación.

Piso. En la puntuación total, el piso ha sido apropiado de acuerdo al criterio de cualificación. Los rendimientos en el año 2004 y 2007 han sido algo superiores a la de los otros años, debido que el piso fue más elevado. Los resultados de la Tabla 55 con respecto al piso sugieren que hay un amplio espacio para hacer discriminaciones de rendimientos bajos, y en que favorecería la descripción del rendimiento.

Tabla 55

Análisis de la gradiente del ítem para el puntaje total a través de los años, sin incluir la subescala Vocabulario

PD	2004			2005			2006			2007		
	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]
78	-	-	-	-	-	-	-	-	-	141	-	-
77	-	-	-	139	-	-	142	-	-	136	6	No
76	142	-	-	133	6	No	136	6	No	133	3	Sí
75	136	6	No	131	2	Sí	135	2	Sí	130	3	Sí
74	131	5	Sí	130	1	Sí	132	2	Sí	128	1	Sí
73	128	3	Sí	127	3	Sí	128	4	Sí	125	3	Sí
72	127	1	Sí	124	2	Sí	126	3	Sí	122	4	Sí
71	127	1	Sí	123	1	Sí	124	2	Sí	121	1	Sí
70	127	0	Sí	123	1	Sí	123	1	Sí	120	0	Sí
69	125	1	Sí	122	1	Sí	122	1	Sí	120	1	Sí
68	123	2	Sí	119	3	Sí	121	1	Sí	119	1	Sí
67	121	2	Sí	116	2	Sí	120	1	Sí	118	1	Sí
66	120	1	Sí	116	1	Sí	119	1	Sí	117	1	Sí
65	119	1	Sí	115	1	Sí	118	1	Sí	115	1	Sí
64	118	1	Sí	115	1	Sí	117	1	Sí	115	1	Sí
63	117	1	Sí	114	1	Sí	117	0	Sí	114	1	Sí
62	116	1	Sí	113	1	Sí	116	1	Sí	112	1	Sí
61	115	1	Sí	112	1	Sí	115	2	Sí	111	2	Sí
60	114	1	Sí	110	1	Sí	114	1	Sí	110	1	Sí
59	113	1	Sí	110	0	Sí	113	0	Sí	109	1	Sí
58	112	1	Sí	110	0	Sí	112	1	Sí	108	1	Sí
57	111	1	Sí	109	1	Sí	111	1	Sí	107	1	Sí
56	110	1	Sí	107	1	Sí	110	1	Sí	107	1	Sí
55	109	1	Sí	106	1	Sí	108	2	Sí	106	1	Sí
54	108	1	Sí	106	1	Sí	107	1	Sí	105	1	Sí
53	106	2	Sí	105	1	Sí	105	1	Sí	103	1	Sí
52	105	1	Sí	104	1	Sí	104	1	Sí	103	1	Sí
51	105	1	Sí	103	1	Sí	103	1	Sí	102	1	Sí
50	104	1	Sí	102	1	Sí	102	1	Sí	101	1	Sí
49	103	1	Sí	101	1	Sí	101	1	Sí	99	1	Sí
48	102	1	Sí	100	1	Sí	101	1	Sí	98	1	Sí
47	101	1	Sí	99	1	Sí	100	1	Sí	97	1	Sí
46	100	1	Sí	99	0	Sí	99	1	Sí	95	1	Sí
45	99	1	Sí	98	1	Sí	98	1	Sí	95	0	Sí
44	98	0	Sí	96	1	Sí	97	1	Sí	94	1	Sí
43	98	0	Sí	96	1	Sí	97	0	Sí	94	1	Sí
42	98	0	Sí	95	0	Sí	96	0	Sí	92	1	Sí
41	96	1	Sí	95	1	Sí	96	1	Sí	92	1	Sí
40	95	1	Sí	94	0	Sí	95	1	Sí	91	1	Sí
39	94	1	Sí	94	0	Sí	94	1	Sí	91	1	Sí
38	93	1	Sí	93	1	Sí	94	0	Sí	89	2	Sí
37	92	1	Sí	93	1	Sí	93	1	Sí	87	2	Sí
36	91	1	Sí	91	1	Sí	92	1	Sí	86	1	Sí

PD	2004			2005			2006			2007		
	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]	PE	Dif.[a]	Cual.[b]
Continua de la página anterior												
35	90	1	Sí	90	1	Sí	91	1	Sí	85	1	Sí
34	90	0	Sí	90	1	Sí	90	1	Sí	84	1	Sí
33	89	1	Sí	88	1	Sí	89	1	Sí	82	1	Sí
32	88	1	Sí	87	1	Sí	89	1	Sí	81	1	Sí
31	87	1	Sí	86	1	Sí	88	1	Sí	81	0	Sí
30	86	1	Sí	85	1	Sí	87	1	Sí	80	1	Sí
29	86	1	Sí	84	0	Sí	86	1	Sí	78	2	Sí
28	85	0	Sí	84	0	Sí	85	1	Sí	77	1	Sí
27	84	1	Sí	84	0	Sí	85	1	Sí	77	0	Sí
26	82	2	Sí	84	0	Sí	84	1	Sí	76	0	Sí
25	81	1	Sí	84	0	Sí	82	2	Sí	75	1	Sí
24	81	0	Sí	84	0	Sí	81	1	Sí	75	1	Sí
23	80	1	Sí	83	1	Sí	80	1	Sí	74	1	Sí
22	79	1	Sí	81	1	Sí	79	1	Sí	72	1	Sí
21	77	1	Sí	81	0	Sí	79	1	Sí	71	1	Sí
20	76	1	Sí	80	1	Sí	78	1	Sí	70	1	Sí
19	75	2	Sí	79	1	Sí	76	1	Sí	70	0	Sí
18	74	1	Sí	77	2	Sí	76	0	Sí	69	2	Sí
17	72	2	Sí	76	1	Sí	75	1	Sí	64	4	Sí
16	70	2	Sí	76	1	Sí	74	1	Sí	62	2	Sí
15	69	1	Sí	75	1	Sí	73	1	Sí	62	0	Sí
14	67	2	Sí	75	0	Sí	72	1	Sí	62	0	Sí
13	65	1	Sí	74	1	Sí	70	1	Sí	62	0	Sí
12	64	2	Sí	73	1	Sí	70	1	Sí	62	0	Sí
11	62	2	Sí	72	1	Sí	70	0	Sí	62	0	Sí
10	62	0	Sí	71	1	Sí	68	2	Sí	62	0	Sí
9	58	3	Sí	70	1	Sí	65	2	Sí	62	0	Sí
8	40	18	No	69	1	Sí	65	0	Sí	59	4	Sí
7	-	-	-	69	0	Sí	65	0	Sí	40	19	No
6	-	-	-	69	0	Sí	64	2	Sí	-	-	-
5	-	-	-	69	0	Sí	62	2	Sí	-	-	-
4	-	-	-	67	2	Sí	62	0	Sí	-	-	-
3	-	-	-	65	2	Sí	62	0	Sí	-	-	-
2	-	-	-	61	4	Sí	62	0	Sí	-	-	-
1	-	-	-	40	21	No	58	3	Sí	-	-	-
0	-	-	-	40	0		40	18	No	-	-	-

[a]Diferencia entre el puntaje correspondiente al resultado de la celda y el puntaje siguiente superior.
[b]Resultado de comparar los dos puntajes consecutivos y el criterio usado (Sí cumple o No cumple con el criterio)

AJUSTE DE LOS PUNTAJES A UNA DISTRIBUCIÓN PSICOMÉTRICA

Luego de evaluar las diferencias entre grupos y no hallarse discrepancias sustanciales en los estadísticos descriptivos, se tomará toda la muestra disponible para generar una distribución que permita derivar normas interpretativas. Para tal fin, se ajustó la distribución empírica de puntajes al modelo beta binomial de cuatro parámetros (4PB) (Tabla 56). Los momentos de la distribución de los puntajes, ajustadas y no ajustadas al modelo 4PB, así como los parámetros 4PB aparecen en la Tabla 56. Estos parámetros para el ajuste corresponden a la forma de la distribución (α y β), el límite inferior (L. inf.) y el límite superior (L. sup.).

Los puntajes con un mayor grado de ajuste fue Letras y Palabras, pues obtuvo la tasa χ^2 /g.l. de menor magnitud comparado con las obtenidas en los otros puntajes. Habilidades Fonológicas y Percepción Visual mostraron un ajuste similar entre sí, y algo mejor comparado con el ajuste de Habilidades Cuantitativas. Excepto el puntaje total, el ajuste de todos los puntajes no superó el valor $p = 0.05$, lo que indica que el ajuste al modelo 4PB fue adecuado. El puntaje total mostró un ajuste poco satisfactorio, aún luego de corregir el criterio de significancia estadística ($p = 0.05$) con el ajuste Bonferroni. Por lo tanto, el ajuste en esta distribución al modelo 4PB no parece apropiada, pero la magnitud de la tasa χ^2 /g.l. es aún es pequeña y cercana al criterio de ajuste aceptable (tasa ≤ 2). Con el ajuste, el puntaje total se hizo algo más curtósico pero su grado de asimetría prácticamente no se modificó.

Tabla 56

Distribución de los puntajes directos y ajustados: Momentos y parámetros

	Let	Fon	Pervis	Cuant	BDPG
Momentos originales (puntaje directo)					
M	8.88	9.04	13.56	15.47	45.97
DE	4.68	4.01	4.72	5.24	15.80
Asim.	0.21	-0.0006	-0.67	-0.76	-0.32
Cur	0.89	2.35	3.00	2.99	2.65
Modelo 4PB					
Momentos ajustados					
M	8.88	9.04	13.56	15.47	45.97
DE	4.23	3.67	4.33	4.23	15.39
Asim.	0.28	0.04	-0.78	-0.87	-0.32
Cur	2.01	2.5	3.36	3.17	2.72
Parámetros					
$.k$	2.62	2.13	0.90	3.16	6.95
$.\alpha$	0.85	1.66	2.04	1.89	2.84
$.\beta$	1.91	1.75	0.79	0.68	1.80
L. inf.	0.13	0.090	0.00	0.00	0.00
L. sup.	1.00	1.00	0.89	0.87	0.97
Prueba de ajuste					
χ^2(g.l.)	10.74 (16)[a]	28.27 (15)[a]	36.05 (19)[a]	25.60 (22)[a]	166.88 (75)**[a]
Razón χ^2 /g.l.	0.67	1.88	1.83	1.16	2.22
N ítems	18	17	21	24	80
KR_{horst}	0.89	0.84	0.84	0.88	0.95

Asim: coeficiente de asimetría. Cur: coeficiente de curtosis. Let. Letras y Palabras. Fon: Habilidades Fonológicas. Pervis: Percepción Visal. Cuant: Habilidads Cuantitativas. BDPG: Puntaje total. [a]corrección Bonferroni aplicado (0.05 / 5 = 0.01). **: $p < 0.001$

Se observa también que el ajuste produjo una leve reducción de la dispersión de los puntajes; esta reducción respecto a sus desviaciones estándares originales estuvieron en un rango del 9% (Percepción Visual) hasta un 23% (Habilidades Cuantitativas). Cuantitativamente, este cambio se vio en algunos lugares decimales y pueden tener algún impacto sobre la curtosis de los puntajes y la gradiente del ítem. Los cambios observados

en la magnitud de la curtosis se incrementó, luego del ajuste al modelo 4PB. Es decir, esta variación estuvo entre 2% (puntaje total BDPG) y 12% (Percepción Visual) menos que las curtosis originales en las mismas variables, además de un gran cambio en el puntaje de Letras y Palabras (125% más curtósico). Esta contradicción parece no coherente pues la disminución de la desviación estándar debería producir una distribución menos curtósica (Cohen, 2001), pero el modelo 4PB contiene un parámetro de confiabilidad del puntaje analizado, y su magnitud podría explicar esta aparente contradicción.

Las frecuencias y probabilidades ajustadas y no ajustadas de presentan en las siguientes tablas: Tabla 57, Tabla 58, Tabla 59, Tabla 60 y Tabla 61. Se observa que generalmente las diferencias entre las frecuencias no ajustadas y ajustadas han estado alrededor de 0, excepto en algunos puntos extremos de la distribución de Percepción Visual (Tabla 59) y Habilidades Cuantitativas (Tabla 60), pero que sin embargo ha tendido a ser muy pequeña. En este último, por ejemplo, el χ^2 para el puntaje 23 ((17 – 9.86)^2 / 9.86 fue 5.15, que es el más elevado en relación al resto de puntajes de Habilidades Cuantitativas, pero es bajo en términos absolutos. Por otro lado, las diferencias en términos de χ^2 en la escala Percepción Visual han sido alrededor de 4 en los puntajes próximos a 1 desviación estándar sobre el promedio.

Observando las distribuciones observadas y ajustadas, las frecuencias observadas (f_{ob}) son algo mayores en las colas de la distribución que las frecuencias ajustadas (f_{aj}). En la zona central, el patrón no ha sido claro. El efecto final del ajuste a la distribución teórica es que se tiende a “suavizar” la continuidad entre los puntajes adyacentes, de tal modo que los cambios de frecuencia entre un puntaje y otro son menos severos.

Los resultados del presente análisis para la elaboración de normas señalas que el ajuste de la distribución de puntajes de las escalas del BDPG a la distribución 4PB es adecuada, y posiblemente más representativa que un ajuste a la distribución normal.

Tabla 57
Conocimiento de Letras y Palabras: Frecuencias directas y ajustadas por el modelo beta binomial compuesto (4 parámetros)

PD[a]	f_{ob}[b]	f_{aj}[c]	P_{ob}[d]	p_{aj}[e]
0	2	1.54428	0.00434	0.00335
1	9	9.41815	0.01952	0.02043
2	27	22.02511	0.05857	0.04778
3	29	31.99770	0.06291	0.06941
4	29	35.82472	0.06291	0.07771
5	37	35.41811	0.08026	0.07683
6	37	33.63092	0.08026	0.07295
7	30	31.85217	0.06508	0.06909
8	25	30.34868	0.05423	0.06583
9	34	29.04324	0.07375	0.06300
10	30	27.84438	0.06508	0.06040
11	34	26.69447	0.07375	0.05791
12	24	25.55348	0.05206	0.05543
13	23	24.38452	0.04989	0.05289
14	18	23.14368	0.03905	0.05020
15	23	21.76601	0.04989	0.04721
16	15	20.13122	0.03254	0.04367
17	21	17.92970	0.04555	0.03889
18	14	12.44945	0.03037	0.02701

[a]PD = Puntaje directo, N total = 461. [b]f_{ob} = Frecuencia observada. [c]f_{aj} = Frecuencia ajustada al modelo 4PB. [d]p_{ob} = Proporción observada. [e]p_{aj} = Proporción ajustada al modelo 4PB

Tabla 58
Habilidades Fonológicas: Frecuencias directas y ajustadas por el modelo beta binomial compuesto (4 parámetros)

PD[a]	f_{ob}[b]	f_{aj}[c]	P_{ob}[d]	p_{aj}[e]
0	5	1.29145	0.01085	0.00280
1	7	6.26210	0.01518	0.01358
2	12	13.84781	0.02603	0.03004
3	14	21.30288	0.03037	0.04621
4	31	27.22835	0.06725	0.05906
5	28	31.56484	0.06074	0.06847
6	32	34.61863	0.06941	0.07509
7	33	36.62725	0.07158	0.07945
8	46	37.72444	0.09978	0.08183
9	42	37.98316	0.09111	0.08239
10	40	37.44207	0.08677	0.08122
11	44	36.11581	0.09544	0.07834
12	36	33.99739	0.07809	0.07375
13	20	31.05544	0.04338	0.06737
14	21	27.22429	0.04555	0.05905
15	19	22.37857	0.04121	0.04854
16	18	16.25691	0.03905	0.03526
17	13	8.07861	0.02820	0.01752

[a]PD = Puntaje directo, N total = 461. [b]f_{ob} = Frecuencia observada. [c]f_{aj} = Frecuencia ajustada al modelo 4PB. [d]p_{ob} = Proporción observada. [e]p_{aj} = Proporción ajustada al modelo 4PB

Tabla 59
Percepción Visual: Frecuencias directas y ajustadas por el modelo beta binomial compuesto (4 parámetros)

PD[a]	f_{ob}[b]	f_{aj}[c]	P_{ob}[d]	p_{aj}[e]
0	4	1.37488	0.00868	0.00298
1	3	2.99216	0.00651	0.00649
2	5	4.68499	0.01085	0.01016
3	8	6.44582	0.01735	0.01398
4	6	8.27457	0.01302	0.01795
5	7	10.17462	0.01518	0.02207
6	8	12.15206	0.01735	0.02636
7	11	14.21564	0.02386	0.03084
8	16	16.37731	0.03471	0.03553
9	16	18.65331	0.03471	0.04046
10	19	21.06600	0.04121	0.04570
11	37	23.64719	0.08026	0.05130
12	29	26.44446	0.06291	0.05736
13	36	29.53289	0.07809	0.06406
14	37	33.02827	0.08026	0.07164
15	45	37.05336	0.09761	0.08038
16	34	41.47640	0.07375	0.08997
17	34	45.13679	0.07375	0.09791
18	30	44.88686	0.06508	0.09737
19	39	36.67172	0.08460	0.07955
20	26	20.82604	0.05640	0.04518
21	11	5.88465	0.02386	0.01276

[a]PD = Puntaje directo, N total = 461. [b]f_{ob} = Frecuencia observada. [c]f_{aj} = Frecuencia ajustada al modelo 4PB. [d]p_{ob} = Proporción observada. [e]p_{aj} = Proporción ajustada al modelo 4PB

Tabla 60
Habilidades Cuantitativas: Frecuencias directas y ajustadas por el modelo beta binomial compuesto (4 parámetros)

PD[a]	f_{ob}[b]	f_{aj}[c]	P_{ob}[d]	p_{aj}[e]
0	3	1.04630	0.00651	0.00227
1	3	2.40823	0.00651	0.00522
2	4	3.72193	0.00868	0.00807
3	2	5.03018	0.00434	0.01091
4	4	6.35190	0.00868	0.01378
5	9	7.70001	0.01952	0.01670
6	11	9.08566	0.02386	0.01971
7	11	10.51993	0.02386	0.02282
8	7	12.01494	0.01518	0.02606
9	14	13.58486	0.03037	0.02947
10	13	15.24707	0.02820	0.03307
11	21	17.02384	0.04555	0.03693
12	18	18.94494	0.03905	0.04110
13	18	21.05270	0.03905	0.04567
14	26	23.41353	0.05640	0.05079
15	32	26.14449	0.06941	0.05671
16	35	29.45641	0.07592	0.06390
17	39	33.64738	0.08460	0.07299
18	34	38.80113	0.07375	0.08417
19	41	43.86772	0.08894	0.09516
20	41	45.59683	0.08894	0.09891
21	27	39.61028	0.05857	0.08592
22	29	25.35759	0.06291	0.05501
23	17	9.86764	0.03688	0.02140
24	2	1.50455	0.00434	0.00326

[a]PD = Puntaje directo, N total = 461. [b]f_{ob} = Frecuencia observada. [c]f_{aj} = Frecuencia ajustada al modelo 4PB. [d]p_{ob} = Proporción observada. [e]p_{aj} = Proporción ajustada al modelo 4PB

Tabla 61

Puntaje Total: Frecuencias directas y ajustadas por el modelo beta binomial compuesto (4 parámetros)

PD[a]	f_{ob}[b]	f_{aj}[c]	p_{ob}[d]	p_{aj}[e]	PD[a]	f_{ob}[b]	f_{aj}[c]	p_{ob}[d]	p_{aj}[e]
0	0	0.02187	0.00000	0.00005	39	5	8.80169	0.01085	0.01909
1	1	0.06957	0.00217	0.00015	40	8	9.00861	0.01735	0.01954
2	1	0.13968	0.00217	0.00030	41	6	9.20452	0.01302	0.01997
3	0	0.23008	0.00000	0.00050	42	9	9.38865	0.01952	0.02037
4	1	0.33916	0.00217	0.00074	43	5	9.56020	0.01085	0.02074
5	0	0.46554	0.00000	0.00101	44	11	9.71835	0.02386	0.02108
6	1	0.60800	0.00217	0.00132	45	12	9.86227	0.02603	0.02139
7	0	0.76546	0.00000	0.00166	46	5	9.99113	0.01085	0.02167
8	1	0.93688	0.00217	0.00203	47	16	10.10406	0.03471	0.02192
9	1	1.12131	0.00217	0.00243	48	11	10.20017	0.02386	0.02213
10	2	1.31783	0.00434	0.00286	49	15	10.27855	0.03254	0.02230
11	1	1.52557	0.00217	0.00331	50	10	10.33828	0.02169	0.02243
12	0	1.74366	0.00000	0.00378	51	12	10.37840	0.02603	0.02251
13	2	1.97130	0.00434	0.00428	52	9	10.39791	0.01952	0.02256
14	1	2.20768	0.00217	0.00479	53	17	10.39577	0.03688	0.02255
15	1	2.45202	0.00217	0.00532	54	12	10.37093	0.02603	0.02250
16	2	2.70355	0.00434	0.00586	55	16	10.32226	0.03471	0.02239
17	3	2.96150	0.00651	0.00642	56	9	10.24859	0.01952	0.02223
18	4	3.22515	0.00868	0.00700	57	12	10.14869	0.02603	0.02201
19	2	3.49374	0.00434	0.00758	58	7	10.02127	0.01518	0.02174
20	5	3.76654	0.01085	0.00817	59	5	9.86495	0.01085	0.02140
21	2	4.04285	0.00434	0.00877	60	6	9.67823	0.01302	0.02099
22	4	4.32192	0.00868	0.00938	61	13	9.45954	0.02820	0.02052
23	4	4.60306	0.00868	0.00998	62	12	9.20714	0.02603	0.01997
24	4	4.88555	0.00868	0.01060	63	5	8.91914	0.01085	0.01935
25	4	5.16868	0.00868	0.01121	64	3	8.59344	0.00651	0.01864
26	7	5.45175	0.01518	0.01183	65	6	8.22766	0.01302	0.01785
27	1	5.73404	0.00217	0.01244	66	8	7.81911	0.01735	0.01696
28	5	6.01485	0.01085	0.01305	67	4	7.36463	0.00868	0.01598
29	4	6.29347	0.00868	0.01365	68	9	6.86045	0.01952	0.01488
30	6	6.56919	0.01302	0.01425	69	6	6.30187	0.01302	0.01367
31	9	6.84131	0.01952	0.01484	70	2	5.68277	0.00434	0.01233
32	4	7.10910	0.00868	0.01542	71	3	4.99473	0.00651	0.01083
33	9	7.37185	0.01952	0.01599	72	7	4.22641	0.01518	0.00917
34	11	7.62884	0.02386	0.01655	73	10	3.36669	0.02169	0.00730
35	5	7.87934	0.01085	0.01709	74	3	2.42155	0.00651	0.00525
36	12	8.12262	0.02603	0.01762	75	2	1.45541	0.00434	0.00316
37	9	8.35793	0.01952	0.01813	76	3	0.63000	0.00651	0.00137
38	14	8.58454	0.03037	0.01862	77	3	0.13900	0.00651	0.00030

[a]PD = Puntaje directo, N total = 461. [b]f_{ob} = Frecuencia observada. [c]f_{aj} = Frecuencia ajustada al modelo 4PB. [d]p_{ob} = Proporción observada. [e]p_{aj} = Proporción ajustada al modelo 4PB

DIFERENCIAS NORMATIVAS

La información expuesta en la Tabla 62, Tabla 63, Tabla 64 y Tabla 65 se refiere a los estadísticos descriptivos utilizados para el presente análisis normativo. Las diferencias normativas se examinaron multivariadamente para probar los efectos de dos variables de posible influencia fija: el género de los participantes y el colegio. El modelo examinado fue completamente factorial, que incluye los efectos principales (género y colegio) y la interacción entre estas variables (género x colegio). Se podría esperar que no habrían diferencias entre los colegios bajo la premisa que la distribución de puntajes en los grupos evaluados no estuviera sesgado por características idiosincrásicas de las condiciones de instrucción preescolar de los niños.

En los años en que se puede hacer la comparación entre colegios, se observa, en general, una leve tendencia de medias mayores en el colegio MI respecto al colegio SM. Las diferencias son, sin embargo, pequeñas. Estas serán evaluadas en los siguientes párrafos.

Tabla 62
Estadísticos descriptivos para el BDPG en el año 2004[a, b]

	MI (n = 96)				SM (n = 93)				Total(n = 189)			
	M	DE	Asim. (Z)	Cur. (Z)	M	DE	Asim. (Z)	Cur. (Z)	M	DE	Asim. (Z)	Cur. (Z)
Let. Pal.	10.21	4.58	-0.15 (-0.40)	-0.88 (-1.37)	7.22	4.86	0.71 (0.46)	-0.44 (-0.94)	8.74	4.94	0.21 (0.04)	-1.03 (-1.3)
Hab. Fon.	10.28	3.76	0.028 (-0.21)	-0.72 (-1.21)	8.42	4.12	0.31 (0.0)	-0.81 (-1.30)	9.37	4.04	0.11 (-0.06)	-0.83 (-1.18)
Per. Visual	12.74	4.19	-0.23 (-0.48)	-0.30 (-0.7)	12.82	4.31	-0.22 (-0.47)	-0.63 (-1.12)	12.78	4.24	-0.22 (-0.40)	-0.49 (-0.84)
Hab. Cuant.	15.21	4.55	-1.07 (-1.32)	0.72 (0.23)	13.28	5.13	-0.41 (-0.66)	-0.74 (-1.23)	14.26	4.93	-0.71 (-0.89)	-0.29 (-0.64)
Hab. Con.	11.19	3.75	-0.29 (-0.53)	-0.94 (-1.43)	10.29	3.84	0.02 (-0.22)	-0.95 (-1.44)	10.75	3.81	-0.13 (-0.31)	-1.0 (-1.35)
BDPG	48.44	14.11	-0.47 (-0.72)	-0.02 (-0.55)	41.73	15.47	0.24 (-0.0)	-0.72 (-1.22)	45.14	15.13	-0.12 (-0.30)	-0.68 (-1.03)

[a] Error estándar de la asimetría: Total = 0.17; Colegio M.I. = 0.24; Colegio S.M. = 0.25
[b] Error estándar de la curtosis: Total = 0.35: Colegio M.I. = 0.48; Colegio S.M. = 0.49

Tabla 63
Estadísticos descriptivos para el BDPG en el año 2005 [a] (n = 107)

	M	DE	Asim. (Z)	Cur. (Z)
Let. Pal.	8.22	4.851	0.44 (1.88)	-0.91 (-1.97)
Hab. Fon.	9.09	4.152	-0.10 (-0.44)	-0.73 (1.57)
Per. Visual	12.78	5.236	-0.71 (-3.03)	-0.13 (-0.27)
Hab. Cuant.	15.99	5.895	-0.79 (-3.38)	-0.16 (0.34)
Hab. Con.	14.14	3.859	-0.98 (-4.19)	1.02 (2.21)
BDPG	65.30	20.483	-0.49 (-2.09)	-0.20 (-0.44)

[a] Error estándar de la asimetría = 0.23; error estándar de la curtosis = 0.46

Tabla 64

Estadísticos descriptivos para el BDPG en entre los años 2006 [a, b]

	MI (n = 94)				SM (n = 93)				Total(n = 187)			
	M	DE	Asim. (Z)	Cur. (Z)	M	DE	Asim. (Z)	Cur. (Z)	M	DE	Asim. (Z)	Cur. (Z)
Let. Pal.	7.83	4.84	0.52 (2.11)	-0.74 (-1.51)	8.86	4.678	0.15 (0.63)	-.092 (-1.87)	8.34	4.779	0.33 (1.8)	-0.90 (-2.55)
Hab. Fon.	8.51	4.34	-0.02 (-0.09)	-0.86 (-1.75)	8.41	3.487	0.10 (0.43)	0.09 (0.19)	8.46	3.928	0.03 (0.17)	-0.49 (-1.38)
Per. Visual	13.47	4.89	-0.68 (-2.76)	0.09 (0.18)	13.44	4.340	-0.40 (-1.64)	-0.51 (-1.04)	13.45	4.616	-0.57 (-3.21)	-0.13 (-0.37)
Hab. Cuant.	15.17	5.85	-0.78 (-3.17)	-0.18 (-0.36)	15.00	5.232	-0.79 (-3.17)	-0.36 (-0.74)	15.09	5.537	-0.78 (-4.4)	-0.25 (-0.71)
Hab. Con.	14.81	3.86	-1.31 (-5.30)	2.28 (1.62)	14.08	2.802	-0.66 (-2.68)	0.07 (0.16)	14.44	3.390	-1.05 (-5.9)	1.79 (5.07)
BDPG	73.17	21.82	-0.61 (-2.48)	0.49 (0.99)	73.20	17.291	-0.17 (-0.69)	-0.81 (-1.65)	73.19	19.647	-0.48 (-2.6)	0.25 (0.71)

Let. Pal: Letras y Palabras. Hab. Fon.: Habilidades Fonológicas. Per. Visual: Percepción Visual. Hab. Cuant.: Habilidad Cuantitativa. Hab. Con.: Habilidad Conceptual. [a] Error estándar de la asimetría: Total = 0.18; Colegio M.I. = 0.25; Colegio S.M. = 0.25. [b] Error estándar de la curtosis: Total = 0.35: Colegio M.I. = 0.49; Colegio S.M. = 0.50

Tabla 65

Estadísticos descriptivos para el BDPG en entre los años 2007 [a, b]

	MI (n = 80)				SM (n = 87)				Total(n = 167)			
	M	DE	Asim. (Z)	Cur. (Z)	M	DE	Asim. (Z)	Cur. (Z)	M	DE	Asim. (Z)	Cur. (Z)
Let. Pal.	10.35	4.47	-0.15 (-0.58)	-0.99 (-1.87)	9.52	4.18	0.23 (0.91)	-0.69 (-1.36)	9.92	4.33	0.04 (0.25)	-0.92 (-2.46)
Hab. Fon.	10.00	4.02	-0.03 (-0.14)	-0.83 (-1.56)	9.37	3.88	0.08 (0.34)	-0.65 (-1.27)	9.67	3.95	0.03 (0.17)	-0.76 (-2.05)
Per. Visual	14.36	5.14	-0.91 (-3.4)	0.08 (0.16)	14.03	3.72	-0.2 (-0.78)	-0.59 (-1.15)	14.19	4.45	-0.68 (-3.66)	0.09 (0.24)
Hab. Cuant.	16.34	4.79	-0.38 (-1.42)	1.45 (2.74)	15.02	4.30	-0.56 (-2.17)	0.31 (0.6)	15.65	4.57	-0.39 (-2.12)	0.92 (2.48)
Hab. Con.	15.16	2.67	-0.5 (-1.86)	-0.29 (-0.55)	14.33	3.78	-0.55 (-2.13)	-0.17 (-0.33)	14.73	3.31	-0.66 (-3.54)	0.21 (0.57)
BDPG	80.16	18.13	-0.53 (-2.04)	0.03 (0.06)	74.08	17.00	0.05 (0.21)	-0.45 (-0.9)	76.99	17.76	-0.2 (-1.15)	-0.4 (-1.1)

Let. Pal: Letras y Palabras. Hab. Fon.: Habilidades Fonológicas. Per. Visual: Percepción Visual. Hab. Cuant.: Habilidad Cuantitativa. Hab. Con.: Habilidad Conceptual. [a]Error estándar de la asimetría: Total = 0.18; Colegio M.I. = 0.26; Colegio S.M. = 0.25. [b]Error estándar de la curtosis: Total = 0.37: Colegio M.I. = 0.53; Colegio S.M. = 0.51.

Como primer paso, se evaluó si el modelo multivariado fue estadísticamente significativo en cada variable dependiente (Tabla 66). Únicamente en el periodo 2004 se hallaron diferencias entre los grupos para las variables dependientes Letras y Palabras (F[3, 188] = 7.31, $p < 0.001$, $\eta^2 = 0.106$), Habilidades Fonológicas (F[3, 188] = 4.51, $p = 0.004$, $\eta^2 = 0.068$) y Habilidades Cuantitativas (F[3, 188] = 2.81, $p = 0.041$, $\eta^2 = 0.04$). Esto indica que sobre estos puntajes, se puede afirmar que hay variabilidad originada en los grupos de participantes. Estas diferencias pueden considerarse entre pequeñas y moderadas (Cohen, 1988; Sink & Stroh, 2006).

Tabla 66
Pruebas multivariadas para la diferencia de medias según género, colegio y la interacción entre ellas

Año	Género	Colegio	Interacción
2004	$\lambda^a = 0.96$ $F(5, 181) = 1.14$ $\eta^2 = 0.03$	$\lambda = 0.86$ $F(5, 181) = 5.85$** $\eta^2 = 0.13$	$\lambda = 0.9$ $F(5, 181) = 0.95$ $\eta^2 = 0.02$
2005	$\lambda^a = 0.97$ $F(5, 101) = 0.49$ $\eta^2 = 0.02$	--	--
2006	$\lambda = 0.98$ $F(4, 180) = 0.4$ $\eta^2 = 0.01$	$\lambda = 0.96$ $F(4, 180) = 1.62$ $\eta^2 = 0.03$	$\lambda = 0.98$ $F(4, 180) = 0.59$ $\eta^2 = 0.01$
2007	$\lambda = 0.95$ $F(4, 160) = 2.01$ $\eta^2 = 0.04$	$\lambda = 0.98$ $F(4, 160) = 0.74$ $\eta^2 = 0.018$	$\lambda = 0.97$ $F(4, 160) = 1.18$ $\eta^2 = 0.02$

[a] λ: Lambda de Wilks. [b] η^2: eta cuadrado parcial. ** $p < 0.01$

El siguiente paso fue determinar qué efectos fueron estadísticamente significativos. Excepto el colegio de procedencia de los niños en solo uno de los periodos de recolección de datos (año 2004), observamos que en los demás años ni la interacción

ni los efectos principales ambas variables tuvieron efectos estadísticamente significativos. Ya que se hallaron diferencias significativas en el grupo del año 2004, se aplicaron las pruebas univariadas para examinar más cercanamente el origen de las diferencias. Se halló que las puntuaciones del colegio MI fueron mayores que las del colegio SM en la escala de Letras y Palabras (F[1, 185] = 21.39, p< 0.001), Habilidades Fonológicas (F[1, 185] = 12.42, p = 0.001), y Habilidades Cuantitativas (F[1, 185] = 8.13, p = 0.005); la magnitud de estas diferencias fueron (en la métrica de η^2), 0.10, 0.06 y 0.042, respectivamente, que están dentro de lo que usualmente se considera una diferencia trivial (Cohen, 1988; Sink & Stroh, 2006). Estos resultados sugieren que los puntajes medios entre los periodos 2005 al 2006 pueden ser considerados similares y, por lo tanto, la muestra total puede servir para hacer baremos; en cambio, la información proveniente del periodo 2004 no se incluirá para obtener normas preliminares de rendimiento en la PDBG

Otras características distribucionales también no mostraron grandes diferencias entre ellas. Aunque no evaluados con pruebas estadísticas inferenciales, la desviación estándar entre los grupos en cada subescala no fueron muy diferentes. Por otro lado, respecto a la simetría distribucional, estas fueron generalmente similares, excepto algunas subescalas, por ejemplo Letras y Palabras (Tabla 64) y Percepción Visual (Tabla 65). La curtosis para las subescalas en los grupos comparados también fue similar, pero algunas mostraron grandes diferencias, por ejemplo en Habilidades Fonológicas y Habilidades Conceptuales (Tabla 64), y Percepción Visual y Habilidades Cuantitativas (Tabla 65). Sin embargo, las mayores diferencias entre los grupos en estas características estadísticas generalmente no se repiten en cada año. El puntaje total ha tendido a mostrar diferencias mayores en comparación a sus subescalas, lo que estaría influenciado por el efecto acumulativo de las variables componentes del puntaje total.

CAPÍTULO 4:
DISCUSIÓN

Evidencias de validez de contenido (Análisis de ítems)

La hipótesis respecto al análisis de ítems cubrió dos aspectos: la invariabilidad y la adecuabilidad. En la primera situación, se probó si los parámetros de dificultad y discriminación son constantes a través de los colegios en cada año (comparación entre-año) y entre los años (comparación intra-año).

Al examinar estas hipótesis sobre los ítems, asumimos en primer lugar que estos cumplirían con los estándares recomendados para influenciar en una adecuada tasa de identificación de niños con problemas potenciales de rendimiento escolar y de describir sus habilidades. Los fundamentos de un adecuado contenido, por lo tanto, descansarían en que a) el grado de dificultad de los ítems sería constante en una muestra que representa una distribución normal del atributo latente, correspondiente a la habilidad medida por los puntajes de cada subescala; b) que el grado de dificultad estaría dentro de un rango sugerido para una óptima descripción de las habilidades del niño, para la detección temprana de potenciales problemas de rendimiento y que contribuiría a una distribución de puntajes apropiada; y c) que sólo una la habilidad subyacente es está vinculada con las variaciones de éxito o fracaso en las respuestas a cada ítem. Por lo tanto análisis de ítems cubre dos aspectos: la invariabilidad y la adecuabilidad.

Algunas de las diferencias entre la dificultad de los ítems fueron estadísticamente significativas, y ello indica que la diferencia en la proporción de aciertos entre una muestra y otra es un efecto sistemático y no asociado al error de muestreo. Pero estas diferencias parecen eventuales pues tendieron a no estar presentes en todas las comparaciones intra-año y entre-año; y la presente eventualidad de estas diferencias estadísticamente significativas fue evaluada también por el tamaño de tales diferencias, es decir, por la magnitud del efecto. Desde un punto práctico, seguidamente se comprobó que la magnitud de estas diferencias fue generalmente baja o cercana a cero, lo que da un respaldo para afirmar la constancia de la dificultad de los ítems entre los colegios, y a

través de los años. La contradicción entre significancia estadística y práctica no es extraño en los análisis cuantitativos que incluyen estos dos criterios (Rosnow & Rosenthal, 2008), pues la significancia estadística está influenciada fuertemente por el tamaño muestral, mientras que la significancia práctica o magnitud del efecto no (Cohen, 2001). Entonces, la evaluación de la hipótesis de igualdad en la dificultad de los ítems puede aceptarse con seguridad, y las diferencias en algunos de los ítems pueden explicarse como eventos aislados y posiblemente no replicables en otros muestreos. Es probable que las variaciones en la administración de los ítems produjeran elevaciones o disminuciones en los desempeños de los niños, y menos a que los ítems favorezcan sistemáticamente a los niños de un colegio.

La pequeña y eventual variabilidad en la dificultad de los ítems también sugiere que el contenido de los ítems no necesariamente dependió de aprendizajes únicos entre los colegios y años, y que además los ítems poseen una potencial generalizabilidad, pudiéndose asumir confiablemente que es apropiado una estimación del parámetro de dificultad basado en el grupo total de participantes.

Una de las observaciones que resaltan es que, aunque las diferencias generalmente no fueron de magnitud, se reconocía una orientación a que los ítems sean más exitosamente respondidos por el colegio MI. Ya que la elaboración y adaptación de los ítems se sustentó en la construcción de una prueba de uso general y no localizado en una sola institución educativa, este patrón no podría originarse de un sesgo en el proceso de elección de los ítems, ni en la administración del instrumento. Parece más razonable conjeturar que los niños provenientes del colegio SM muestran menos habilidades y conocimientos para responder a los ítems, y que los antecedentes sociales y educativos o un proceso de auto-selección en la matrícula en este colegio podrían congregar a niños con menos competencias. Estas diferencias potenciales podrían detectarse en un análisis

cualitativo o directo sobre la calidad de la estimulación recibida en el hogar y en los centros preescolares de donde provenían. Esta hipótesis es apenas apoyada por las características demográficas de las familias de los niños. En la Tabla 5, la distribución del nivel educativo del padre de familia entrevistado mostró leves diferencias distribucionales. Es decir, había un mayor porcentaje de padres del colegio Sm que tenían instrucción básica incompleta, mientras que los padres del colegio MI tenían una mayor proporción de padres con estudios más allá del nivel básico. Pero como se mencionó, estas diferencias fueron leves, pero contribuir a explicar las diferencias en el parámetro de dificultad de los ítems. Hay que considerar que una de las variables sociales observada para diferenciar los niveles socioeconómicos y de desarrollo humano es el nivel educativo alcanzado por el jefe de la familia (FONCODES, 2006). Considerando el modelo de medición y los propósitos del BDPG, la dificultad de los ítems hallado de casi todos los ítems favorecen la discriminación de niños en el rango medio y bajo de dificultad (Erford & Biddison, 2006)

En otra de las dimensiones la evaluación psicométrica de los ítems, es decir la discriminación, se obtuvo resultados que pueden ser considerados en general buenos. Esta valoración se sustenta en la magnitud promedio de la correlación ítem-test en la muestra total y en las submuestras y en la pequeña variabilidad de la magnitud de esta correlación entre los periodos de evaluación (análisis entre-años) y entre los colegios (análisis intra-año). Estos dos criterios se cumplieron de un modo que se considera satisfactorio, pues la tendencia general en los ítems de las subescalas fue el cumplimiento o su cercana aproximación con estos criterios. Como las pruebas estadísticas aplicadas (χ^2 para varias correlaciones independientes, y Z para dos correlaciones independientes) mostraron que no se rechaza la hipótesis nula de igualdad entre ellas, entonces se puede asumir que las diferencias halladas provienen más bien de variaciones aleatorias y no sistemáticas en las

poblaciones de origen. Sin embargo, esta conclusión debe ser moderada por interpretaciones complementarias a la aceptación de la hipótesis nula. Es decir, que la hipótesis nula para evaluar las diferencias en la discriminación fue de igualdad en la magnitud de las correlaciones en la población de interés o que las diferencias entre ellas serán cero (Chen & Popovich, 2002), pero es razonable definir que esta hipótesis es considerada falsa desde su planteamiento (Waller, 2004) y que habrán diferencias diferentes de cero que puedan o no ser estadísticamente significativas. Por lo tanto, los resultados realmente no suponen que la propiedad discriminativa de los ítems en los colegios y los periodos de evaluación muestreados sea igual, sino que no son lo suficientemente variables como para interpretarlos separadamente. Aunque cada ítem aportará desigualmente a la variabilidad de su subescala, hay evidencias que pueden ser consideradas similares en las condiciones en que ocurre la evaluación grupal de niños, especialmente en la práctica profesional. Con estos resultados, también se puede plantear que los factores irrelevantes al constructo en las pruebas parece ser relativamente constante, y que la habilidad muestreada por el ítem podrá estar menos "contaminado" por influencias aleatorias sobre el rendimiento total operacionalizado en el puntaje escalar.

Aunque el nivel general de la discriminación del ítem fue aceptable, se observaron algunas variaciones en la subescala Vocabulario/Conceptos que sugerirían una inestabilidad discriminatoria no satisfactoria. Se halló que en cada periodo de evaluación, las magnitudes eran más heterogéneas que en las demás subescalas y que hubo tendencias claras en que los niños de un colegio tendían a puntuar más. Esta variabilidad puede asociarse a diferencias en la cantidad y calidad de la comprensión verbal de los niños evaluados, a la estimulación en el hogar y particularmente a los centros preescolares de origen (Merino & Muñoz, 2007). Pero en general, estas diferencias las consideramos

insatisfactorias para un instrumento que pretende ser sensible a los aprendizajes preparatorios de la lectura y matemáticas. Es probable que el monto de aprendizajes formales (instrucción preescolar) y no formales (hogar) que reciben los niños antes de la evaluación haya tenido un mayor impacto en los puntajes de Vocabulario/Conceptos, especialmente vinculados con el nivel socioeconómico, la instrucción de la familia y la estimulación en el hogar asociadas a estos aspectos demográficos (Basilio, Puccini, Silva, & Pedromonico, 2005; Merino & Muñoz. 2007). El factor verbal puede ser una importante variable en el rendimiento escolar, pero está más involucrado en edades más avanzadas, incluso desde el 2do grado cuando las habilidades perceptuales son necesarias pero ya no suficientes (Solan, Mozlin & Rumpf, 1985). Todos estos aspectos empíricos, racionales y derivados de la literatura dan respaldo para remover este puntaje del BDPG.

Exceptuando la subescala Conceptos/Vocabulario, ni en las comparaciones intra-año y entre-año se pudo observar un patrón sistemático de cambios entre los coeficientes de dificultad y discriminación. Así, en las comparaciones intra-año, en algunas escalas los coeficientes eran relativamente más elevados en un colegio, mientras que en otras escalas ocurría lo contrario o eran apenas diferentes. También, algunos ítems mostraban diferencias estadísticamente significativas entre los colegios dentro de un mismo año, pero ello no ocurría en otro año. En las comparaciones entre-años, también no se pudo detectar cambios sistemáticos. Y aun cuando hubo diferencias no-cero, estas no parecieron ser de gran magnitud como para sugerir un efecto único e importante producido en el periodo de evaluación específico.

El resultado final del análisis de los ítems fue que los parámetros de dificultad, discriminación y estabilidad entre-grupos de las subescalas, exceptuando Vocabulario/Conceptos, presentaron criterios satisfactorios para aceptarlos e integrarlos

en una unidad coherente, con una estructura interna estable, y cuyos ítems contribuyen favorablemente en la definición del constructo que ellos representan.

Respecto a la homogeneidad, en conjunto los ítems mostraron un rango aceptable de homogeneidad en la mayoría de las escalas, y de acuerdo a la magnitud de estas correlaciones inter-ítem, las tareas estarían representando un constructo de amplio rango. De acuerdo a las sugerencias propuestas por Briggs y Cheek (1986) y Clark y Watson (1995), valores menos que 0.15 indican heterogeneidad en los ítems y posiblemente multidimensionalidad (Piedmont & Hyland, 1993), y alrededor o más de este valor un nivel de homogeneidad esperable para representar constructos de amplio rango, generales y menos específicos. Los resultados de la homogeneidad no son elevados, pero es esperable que las tareas de las escalas contengan un grado moderado de heterogeneidad entre ellas, ya que se diseñaron para muestrear una dimensión que puede ser presentada en varios tipos de tareas. Los niveles de homogeneidad hallados aún son algo bajos en comparación a otras escalas estandarizadas para habilidades pre-académicas que tienen como respaldado por su elevada unidimensionalidad (Erford, 2004; Erford & Biddison, 2006), pero pueden considerarse aceptablemente homogéneas tomando en cuenta los resultados de la confiabilidad y la dimensionalidad.

Los resultados de la homogeneidad revelaron también un leve patrón de mayores magnitudes en las correlaciones inter-item en una de las muestras. Aunque las diferencias fueron pequeñas como para señalar un efecto diferencial del contexto de procedencia de los niños, esta variación puede advertir la inclusión de factores irrelevantes al constructo que impactaron a un grupo y no a otro. Considerando que el contenido de las escalas se seleccionó para tener una amplia aplicación de grupo y en varias poblaciones, podría esperarse una variación en el nivel de desempeño por bajos niveles de aprendizajes pero no por variaciones en las relaciones internas entre las tareas. Ya que las tareas descansan

sobre el presupuesto de moderada homogeneidad, las potenciales variaciones en la homogeneidad pueden considerarse más bien eventuales.

Sin embargo, esta conclusión no parece aplicarse con la escala Vocabulario/Conceptos, pues la magnitudes de las correlaciones inter-ítems fueron más bajas que el resto, y es un resultado especialmente consistente entre las muestras intra-año y entre-año. Aunque el contenido de esta escala pueda ser considerado importante en la predicción del rendimiento académico posterior, no parece adecuarse técnicamente al modelo planteado inicialmente para la BDPG. Finalmente, ya que se considera que la homogeneidad de los ítems es un aspecto psicométrico relacionado con la unidimensionalidad (Piedmont & Hyland, 1993), los resultados de este aspecto psicométrico no favorecen la potencial unidimensionalidad en cada escala.

Evidencias de la dimensionalidad

Los resultados indicaron que algunos de los índices de ajustes fueron bajos, mientras paralelamente otros mostraron que el modelo evaluado tenía buen ajuste. Plausiblemente, la falta de modelación de las covariaciones de algunas de las parcelas (clústeres) dentro de un factor puede influenciar en la obtención de bajos índices de ajuste (Hall, Snell, & Foust, 1999). Dado que las parcelas formadas se hicieron sobre una base racional y no empírica, los ítems agrupados tendrían más coherencia teórica con el proceso de re-construcción del BDPG. Sin embargo, este método de parcelamiento también capitalizaría las diferencias en las propiedades estadísticas univariadas de los ítems, y en conjunto estos mostrarían fuerte asimetría y exceso de curtosis. Esto fue justamente lo que se halló en la gran mayoría de las parcelas. Con la ventaja de formar parcelas también vienen algunas desventajas, entre ellas, que no se obtienen los parámetros para los ítems y pueden existir malas especificaciones (Hall et al., 1999;

Little, Cunningham, Shahar, & Widaman, 2002), y que sus relaciones pueden alterarse en otra muestra. Afortunadamente, las correlaciones convergentes y divergentes entre las parcelas, y los indicadores de ajuste, dieron respaldo favorable a cada una. En dos de las parcelas de la escala de habilidad perceptual visual se requirió liberar los términos de error de dos ítems, pero esto no fue sorpresas pues ambas son parte de la misma tarea. Esto indicaría que esta re-especificación también sería necesaria en el modelamiento de todo el instrumento.

Los resultados de la evaluación de la dimensionalidad revelaron que una dimensión de alto orden, y los factores de primer orden parecen ser la que mejor representación estructural del instrumento. Esto permite simplificar la interpretación y explicar coherentemente las relaciones entre los contenidos (Chen, Sousa, & West, 2005), además de justificar empíricamente la formación de un puntaje total.

Los peores ajustes ocurrieron con los modelos de ortogonalidad y oblicuidad extremos, y por lo tanto no hay evidencias de independencia o redundancia conceptual entre las mismas. Esto añade una perspectiva optimista sobre el BDPG respecto a la validez discriminativa entre sus puntajes, y la varianza común subyacente a las mismas. Sin embargo, las correlaciones altas entre los factores realzaron tres cosas: primero, que las habilidades muestreadas deben comprenderse mejor cuando interaccionan entre sí; segundo, que las fuertes relaciones no suponen redundancia o pérdida de validez discriminativa entre ellas; y finalmente, es razonable presunción de un factor general que permita una parsimoniosa interpretación de los puntajes. La propuesta de un factor general es más ventajosa cuando se requiere describir el desempeño total del niño con un solo indicador, además que intrínsecamente muestra la mayor confiabilidad. Usando las cargas factoriales de las parcelas, la confiabilidad fue definitivamente alta para el factor de segundo orden. Este nivel de confiabilidad permite que sea pertinente para propósitos

clínicos, situación en que se exige muy bajos niveles de error de medición debido a las consecuencias críticas sobre el examinado (Charter, 2003; Nunnally & Bernstein, 1995).

Evidencias de validez

Validez de constructo: Correlaciones convergentes y divergentes.

La covariación convergente y divergente entre los puntajes de las subescalas del BDPG y las pruebas de validación (Prueba de Diagnóstico Preescolar, PDP; y Test de Desarrollo Psicomotor, TEPSI) ha ido parcialmente en la dirección de las hipótesis. El puntaje total del BDPG covaría en un monto elevado con la Prueba de Diagnóstico Preescolar (PDP), de tal modo que el ordenamiento de los niños evaluados por ambos instrumentos tiende a ser altamente similar. Esto sugiere un alto monto de varianza común entre ellos. Esto pondría en cuestionamiento al puntaje total del BDPG, pues si es aparente que los procesos medidos en la PDP pueden comparten similaridad con la habilidad global medida por el BDPG, la magnitud de estas habilidades compartidas no debería ser alta sino más bien moderada o menos, pues la BDPG pretende añadir varianza explicativa del desempeño pre-académico del niño, más allá de otras habilidades evaluadas. Los presentes resultados sugieren que el uso del puntaje total del BDPG podría llevar a resultados similares si se usara una prueba de aptitudes general como el PDP. Ya que la prueba estadística de Braden (1986) no rechazó la hipótesis de convergencia entre ambas, el puntaje total del BDPG podría ser un indicador redundante para evaluar la capacidad cognitiva general del niño cuando se lo adjunta a una medida de aptitudes.

Debido a la alta correlación entre los puntajes totales del PDP y BDPG, parecería existir un posible problema de validez de constructo del puntaje total del BDPG, pues se pondría en cuestionamiento su interpretación. Sin embargo, se debe señalar que el BDPG se creó como una medida de habilidades específicas y no de aptitudes. En segundo lugar,

se debería replicar este estudio en una muestra de otro contexto socioeconómico, y de mayor tamaño muestral. Es posible, específicamente en los niños de esta muestra (PRONOEI), que el grado de familiaridad (o falta de la misma) con el formato de respuesta sea un factor que puede añadir varianza irrelevante al constructo (Nunnally & Beernstein, 1995) e inflar las correlaciones de manera espuria. En otras palabras, ya que ambas pruebas (BDPG y PDP) son similares en usar el formato de opción múltiple para responder, el efecto de la varianza del método podría haber añadido un porcentaje de covariación espuria entre ambas medidas. Efectivamente, el BDPG tiene un formato de opción múltiple, igual a una de las pruebas de aptitudes aplicada (PDP), y por lo tanto parte de la varianza común entre ambas puede ser explicada por el método común a ambas. Esto puede ocasionar un espurio aumento de las correlaciones entre los puntajes, enmascarando una estimación más precisa de la validez. Se requeriría separar esta fuente de covariación en una propuesta experimental, y que por ahora se deja como pregunta abierta a ser respondida en el futuro.

Aunque las correlaciones han tendido a ser elevadas cuando son comparadas con las recomendaciones cualitativas de la magnitud de las correlaciones (Cohen, 1992), hay diferencias cuantitativas y cualitativas que deben enfatizarse. Cuantitativamente, las correlaciones no son lo suficientemente altas como para sustentar que los puntajes como mediciones intercambiables de un constructo relacionado. El respaldo de esta conclusión de halló aplicando un procedimiento que tiene como hipótesis nula la más alta correlación posible entre dos medidas convergentes o divergentes. Esta correlación más alta posible se obtuvo luego de desatenuarla por error de medición (Braden, 1986). Los resultados rechazaron esta hipótesis estadística, y por ello se puede suponer que no existe un completo y serio traslapamiento de los atributos medidos. Por otro lado, las correlaciones elevadas indican que el ordenamiento de la habilidad de los niños será moderadamente

coincidente pero no igual; y la similaridad ocurrida puede explicarse más bien por habilidades que concurren en el momento de dar las respuestas. Se puede señalar que las habilidades medidas por dos escalas diferentes compartan contenidos aprendidos que realmente los vinculan, pero estas no son relevantes a la medición de constructo de interés.

Por otro lado, cualitativamente, se debe de recalcar dos puntos. Primero, que las tareas muestrean dominios de contenido diferentes y por lo tanto, la elevada variabilidad común puede ser una consecuencia espuria del método de evaluación común a ambas (ver párrafos anteriores). En segundo lugar, se puede reconocer un patrón correlacional que sugiere que las funciones cognitivas evaluadas convergen con otras medidas de contenido similar, y que divergen moderadamente con otras. Por ejemplo, se halló que Letras/Palabras covaría más con la comprensión verbal y la aptitud perceptual-visual; y parece razonable este patrón porque el conocimiento de las letras y palabras puede estar asociado con contenidos sociales y académicos en que ocurre una alta demandas de las habilidades de percepción visual, pues esta es la base fundamental para adquirir los aprendizajes en las edades tempranas. Por otro lado, la subescala Habilidades Fonológicas comparte más varianza con los puntajes de percepción visual del PDP, y hay evidencias que sugieren la conexión entre ambas habilidades (Edford, 2004; Edford & Biddison, 2006). El reconocimiento automático de palabras es un componente importante de la evaluación temprana, pues está asociado a la fluidez y comprensión de lectura incluso hasta el 3er grado, y es un predictor de la lectura después de los primeros grados, a diferencia de la habilidad fonológica que impacta más en los grados iniciales de instrucción (Badian, 2001).

Se concluye que el BDPG, aplicado en niños de PRONOEI, se diferencia de pruebas de desarrollo psicomotor general. Pero su diferencia con una prueba de aptitudes

generales para el aprendizaje preescolar sólo se logra parcialmente. Su convergencia con esta misma medida de aptitudes es moderada, pero sus relaciones divergentes han sido mejor sustentadas. En general, las evidencias de constructo basadas en las relaciones convergentes/divergentes solo son moderadamente satisfactorias, y el desempeño en el BDPG tiene una moderada covariación con las aptitudes generales.

Los resultados de la asociación entre el BDPG y una prueba de aptitudes preacadémicas (PDP) y de desarrollo psicomotor (TEPSI) han sido parcialmente aceptables, especialmente con la primera prueba. La covariación con la prueba de desarrollo psicomotor fue adecuada para propósitos de validez divergente, dado que las correlaciones tendieron a ser de baja magnitud, y se presuponía que el BDPG no debería traslaparse con las habilidades de amplio rango evaluadas por el TEPSI. Sin embargo, las habilidades medidas por el BDPG correlacionaron significativamente, y ello indica que comparten alguna varianza común. El enfoque correlacional de Braden (1986) confirmó claramente la no equivalencia cuantitativa entre estos instrumentos. Aún con este patrón de correlaciones bajas, hubo un perfil de covariación esperable entre las subescalas de ambas pruebas. Por ejemplo, ninguna subescala BDPG se asoció linealmente con Motricidad del TEPSI; y Hab. de Percepción Visual tuvo mayor asociación con Coordinación. Esta situación da soporte a la validez de constructo.

Con la prueba de aptitudes preescolares los resultados no fueron tan favorables al BDPG como una medida independiente de estas capacidades generales, específicamente para el puntaje total. De los resultados obtenidos, la elevada correlación entre los puntajes totales y el enfoque de Braden (1986) sugieren un posible traslape de los constructos evaluados. Pero estas correlaciones se hicieron en una muestra de niños de PRONOEIs, y la interpretación debe ser moderada por esta característica muestral. Es posible que la evaluación de las habilidades preacadémicas del aprestamiento se involucre con

capacidades más generales, como el dominio perceptual, la comprensión verbal, conceptos cuantitativos y atención a instrucciones; y estas capacidades de fondo ponen los límites del máximo desempeño posible en los niños. El impacto de estas habilidades fundamentales para afrontar las tareas presentadas ha sido reportado en el contexto de la evaluación preescolar (Merino, 2008; Merino & Muñoz, 2007), en que se sugiere que el factor de comprensión verbal puede añadir varianza irrelevante al constructo en la evaluación de niños preescolares de zonas urbano-marginales.

Aunque las correlaciones fueron elevadas con las subpruebas de aptitudes, usando el enfoque de Braden (1986), sin embargo, estas correlaciones no sugieren una identidad cuantitativa con estas aptitudes, pues se rechazó en cada uno que estas correlaciones fueran similares a la máxima correlación posible. Ya que la muestra es pequeña, es razonable dudar sobre la distribución normal de los puntajes para usar pruebas paramétricas como la *r* de Pearson (Cohen, 2001), pero se asume que las distribuciones de las variables no fueron afectadas severamente por la desviación de la normalidad de algunas de las variables, ya que aún en distribuciones reconocidamente diferentes a la normal, el error de Tipo I tiende a ser mínimamente afectada (Levy, 1977).

Con respecto a las pruebas de habilidad intelectual y el BDPG, se aplicaron dos pruebas consideradas equivalentes respecto al constructo medido: el TONI-2 (Brown et al., 2000) y el DAP:IQ (Reynolds & Hickman, 2004). Los resultados mostraron correlaciones homogéneas entre las muestras analizadas (periodos 2007 y 2006) y entre las pruebas usadas. Las diferencias entre las correlaciones pueden considerarse originadas por error de muestreo, pues no resultaron estadísticamente significativas. Considerando estos resultados correlaciones con el TONI-2 y DAP: IQ, la varianza explicada entre el BDPG y estas medidas de habilidad intelectual estuvieron alrededor de 16% ($r = 0.40$). La magnitud moderada hallada sugiere varios aspectos. Primero, que el aporte de la

habilidad intelectual no verbal debe tomarse en cuenta cuando se trata de cuantificar o explicar la relación entre habilidades cognitivas relacionadas con los aprendizajes pre-académicos.

Esto significa que las medidas de habilidad intelectual pueden explicar una parte importante de la varianza entre el BDPG y otras medidas estandarizadas, y tendría valor metodológico incluir las estimaciones de habilidad intelectual como una variable que se deba parcializar o controlarse estadísticamente. En segundo lugar, las habilidades intelectuales no verbales pueden tener validez incremental en un modelo predictivo del rendimiento durante los primeros años de escolaridad si se lo incluyera en una batería de despistaje. Las medidas de habilidad intelectual aplicadas aquí tienen su mejor uso para los fines de despistaje (Brown et al., 2000; Reynolds & Hickman, 2004), pero no podrían ser consideradas como determinantes para detectar tempranamente problemas en el rendimiento escolar como los que pretende evaluar el BDPG. La relación entre el BDPG y las medidas de habilidad intelectual sugiere además que ciertas operaciones cognitivas son compartidas con ambas. La carga de procesamiento visual y razonamiento con materiales no verbalizados pueden tener un significativo impacto en la comprensión de las tareas del BDPG, así como en su resolución. Aunque el éxito en el BDPG requiere de aprendizajes verbales previos, una parte de ella requiere habilidades en que la expresión verbal es reducida. Estas habilidades pueden estar relacionadas con específicas aptitudes propias de la inteligencia fluida (Thorne & Schaefer, 2004) evaluadas por el TONI-2 y DAP: IQ, y una elevada correlación con ellas podría poner en cuestionamiento el BDPG como una medida fundamentalmente de habilidades pre-académicas. Afortunadamente, la magnitud de las correlaciones puede identificar a estas medidas como divergentes.

Finalmente, cuando se exploraron las relaciones entre el BDPG y las medidas de habilidad visomotora, se obtuvieron correlaciones estadísticamente significativas con

todas las subescalas y con el puntaje total del BDPG. Pero se detectó también un patrón esperado: que las correlaciones con el Bender-QSS fueron siempre más elevadas con el VMI-4. La dirección de estas diferencias de validez entre el Bender-QSS y otras medidas de habilidades visomotoras ha sido reportada en otros estudios (Brannigan & Brunner, 2002; Chan, 2000, 2002; Merino, 2010)

Validez para diferenciar grupos

La BDPG demostró un mayor poder discriminativo frente a otros instrumentos, y uno de los motivos puede provenir en que estos otros instrumentos caracterizan habilidades específicas indirectamente vinculadas con los niños aprendizajes de los niños. Las medidas de percepción visual, habilidad visomotora, inteligencia y preparación perceptual-motor del niño fueron evaluados por tareas que representan estos constructos en algún grado de homogeneidad, pero que son más bien generales respecto al aprendizaje específico de, por ejemplo, las letras, conceptos cuantitativos o discriminación fonológica. Aunque algunas diferencias están más allá del error de muestreo, la interpretación más útil es respecto a la significancia práctica de estas diferencias; por tal motivo, las correlaciones y las diferencias estandarizadas estimadas relevaron patrones interesantes. Por ejemplo, la diferenciación del PGB estuvo alrededor de la capacidad discriminativa del BDPG; y ligeramente menor, la prueba de percepción visual y de integración visomotora. La prueba de Conceptos/Vocabulario solo tuvo diferencias triviales entre los grupos, lo que ayuda a reafirmar su limitada utilidad, un aspecto que ya se había detectado en los otros análisis del presente estudio.

Varios puntajes del BDPG fueron más sensibles a las diferencias, especialmente las que tienen relación con los aprendizajes básicos para la lectura y las matemáticas, con lo que se puede hallar un valor de práctico de este hallazgo. Por ejemplo, los cambios en el rendimiento que son descubiertos por la profesora pueden corresponder con cambios

en los puntajes del BDPG, pues el criterio de la profesora parece ser sensible cuando se le solicita su impresión general sobre áreas específicas de rendimiento en sus alumnos. Sus grupos extremos elegidos pueden mostrar diferencias en los contenidos del BDPG, y el desempeño en esta prueba puede asociarse con la elección de la profesora cuando trata de diferencias grupos de rendimiento extremo.

Se debe señalar que la manera en que se recolectaron los datos determina en parte la variabilidad de los resultados hallados. En otras palabras, debido que la profesora recibió la instrucción de selección de niños con alto y bajo rendimiento, esta consigna debió capturar las diferencias observadas respecto a los aprendizajes que curricularmente ocurren en el aula de clases; por lo tanto, la observación de la profesora válidamente capturó el mejor y peor desempeño en aspectos específicos, pero no ocurrió así con las habilidades de influencia indirecta en los aprendizajes. Si se le hubieran dado consignas que dirijan la atención hacia comportamientos asociados a la habilidad visomotora, la percepción visual, la inteligencia general o el estatus general de aprendizaje, es plausible que las diferencias entre los grupos sean menores en la BDPG y relativamente mayores en los otros constructos divergentes. Por otro lado, el tamaño muestral de los grupos comparados fue pequeño, y el poder estadístico de las pruebas inferenciales se reduce en una relación inversamente proporcional (Cohen, 2001). Es de esperarse que con un mayor tamaño muestral, las diferencias actualmente detectadas como estadísticamente no significativas puedan estar debajo del nivel alfa (0.05).

El método de usar grupos extremos tiene limitaciones respecto a tu poder estadístico, así como en la confiabilidad de las medidas y mantener el presupuesto de linealidad entre los puntajes (Preacher et al., 2005). En los presentes se ha asumido que una relación lineal será subyacente a las asociaciones entre las variables analizadas. No habría motivos para suponer otro tipo de relación entre medidas cognitivas, y que son

fundamentalmente dependientes de la edad. La maduración de las funciones cognitivas avanza con la edad, y la ausencia de una relación lineal entre ellas podría más probablemente representar una ausencia de relación ($r = 0$), no solo lineal sino también no lineal. También se asumieron otras condiciones que son razonables en la evaluación de puntajes del rendimiento cognitivo: por ejemplo, se asumió una razonable aproximación de la distribución de los puntajes hacia la distribución normal teórica, y un aceptable monto de error de medición; esto último fue corroborado con la estimaciones de confiabilidad reportadas en estos datos, tal como se recomienda en la práctica actual de reporte de resultados (AERA et al., 1999). Sin embargo, una replicación de los resultados, mediante la misma estrategia en un grupo mayor de participantes permitirá obtener parámetros más estables. Aunque el uso de un tamaño muestral igual en los extremos atenúa levemente la disminución del poder estadístico (Fowler, 1992), otros efectos como la inestabilidad de la dispersión en los grupos extremos, pueden producir grandes e inesperados cambios sobre las correlaciones estimadas en este método (Preacher et al., 2005), y este es común en situaciones en que se usa pequeñas muestras como este tipo de diseños (Flowler, 1992).

La eficiencia de este método no puede rechazarse completamente, pues puede ser la única fuente de acercamiento de una expresión conductual que se presentaría en una conocida característica. En otras palabras, la variabilidad en los constructos cognitivos puede ser confiablemente explicada con las diferencias en el rendimiento académico, pues este desempeño tiene un vínculo con el procesamiento y habilidades cognitivas. El método aplicado en este estudio es, por lo tanto, justificable teóricamente, y también pragmáticamente. Para esta última situación, el tiempo y los recursos materiales, de tiempo y de personal puede poner a la metodología de grupos extremos como la mejor opción posible (Preacher et al., 2005).

Evidencias de validez concurrente

La asociación lineal entre la BDPG y la Prueba Gestáltiva Antonn Brenner (PGAB) fue examinada en dos grupos, cuyo tamaño muestral fue sustancialmente diferente. Esta situación altamente desafiante para la replicabilidad de los hallazgos correlaciónales, debido a la mayor influencia del error muestral. Sin embargo, las correlaciones encontradas fueron satisfactorias pues fueron estadísticamente similares en ambos grupos. Todas las corrrelaciones fueron de dirección positivas y por lo menos moderadas, resultados que son razonables dado que los dos instrumentos evalúan la preparación del niño para el primer grado (aunque desde distintas concepciones), la BDPG contiene habilidades interrelacionadas y que la PGAB es una medida integrativa de habilidades cognitivas no verbales. Por lo tanto, en conjunto ambas medidas se orientan hacia el mismo conjunto de habilidad preparatoria general del niño para ingresar a primer grado. Las correlaciones no podrían ser más altas pues conceptualmente ambos medidas parten de dos concepciones distintas, pues mientras que la BDPG se orienta hacia las habilidades influencias por la instrucción previa, la PGAB descansa sobre aspectos perceptuales, un componente más fuertemente biológico y madurativo.

Por otro lado, en la observación de las correlaciones entre habilidades específicas, el patrón de covariaciones fue fuerte y consistente entre medidas que involucran directamente habilidades cognitivas no verbales, específicamente la percepción visual y las habilidades cuantitativas del BDPG y el puntaje del PAGB. Est es congruente con la medición esencial de la PAGB, es sus tareas contienen un fuerte componente cuantitativo y perceptual-visual. Estos resultados parecen señalar que si requiere una medida global y rápida de la preparación del niño, la PAGB puede ser una medida suficiente; sin embargo, su aplicación podría limitarse como un componente suplementario a la evaluación de

contenidos más relacionados con la instrucción, como ocurre con la BDPG. Adicionalmente, la BDPG permite una diferenciación de contenidos relativamente independientes, aspecto que no permite el PAGB.

Evidencias de la confiabilidad

La estimación del error de medición de los puntajes es crítico en la estimación de habilidad. En el presente estudio, se exploró la confiabilidad por el método de la consistencia interna. Los resultados señalan ciertos patrones en el modo en que la consistencia interna ha variado entre los colegios (análisis intra-año) y entre los periodos de tiempo (análisis entre-años).

Varias de las diferencias halladas entre los colegios pueden asumirse como efectos del error de muestreo, y por lo tanto se puede aceptar que tales variaciones ocurrieron eventualmente, y de poca influencia en concluir que existe sesgo en el contenido muestreado (Reynolds, 2000). Estas variaciones provienen de varias fuentes reconocidas en la literatura (una excelente presentación clásica puede leerse en Stanley, 1971), y que aditivamente impactan en el error aleatorio en los puntajes del BDPG, pero los resultados presentados sugieren un impacto menor. Las variaciones en la consistencia interna entre los colegios parecieron originarse por la interacción de las condiciones organísmicas de los niños, del ambiente de evaluación y del mismo proceso de evaluación conducido por los examinadores. En varias ocasiones, el ambiente de evaluación fue alterado por excesivo ruido, indisposición momentánea de uno o dos niños simultáneamente, interrupciones de padres de familia, incremento del ruido externo al aula de evaluación y variaciones en la instrucción de las tareas de la prueba.

Aunque este último problema se controló repetidamente por el monitoreo coordinado entre los miembros del equipo, su interacción con la disponibilidad interna

del niño no se puede conocer con exactitud. La acumulación de estos efectos únicos en la respuesta del niño hizo que las respuestas puedan ser menos consistentes como para obtener coeficientes elevados, pero los coeficientes de consistencia interna solo reflejan una imagen estática y "fotográfica" (Feldt & Brennan, 1989) de estos efectos. Por lo tanto, se pueden obtener mejores niveles de consistencia interna si las condiciones de administración se mejoran, tomando en cuenta que la aplicación grupal tiende capitalizar la influencia aleatoria de varias fuentes de error. Una la aplicación individualizada del BDPG podría superar estos inconvenientes.

Los resultados de la consistencia interna y su comparación entre-años e intra-años sugieren que la magnitud del error de medición es aceptable y moderadamente constante. Aunque esta propiedad interacciona con la dispersión de la muestra (Feldt & Brennan, 1989) y la homogeneidad de los ítems (Piedmont & Hyland, 1993), podemos concluir que en una aplicación estandarizada en que ocurren leves variaciones en las instrucciones y factores aleatorios de impacto leve, se pueden obtener rangos aceptables de consistencia interna que garantizan respuestas moderadamente homogéneas a los ítems. Ya que la aplicación del BDPG se hizo grupalmente y ello deja una puerta abierta a más errores aleatorios, el nivel hallado de la consistencia interna puede ser considerado muy favorable en el contexto de su aplicación en la detección temprana de niños con potenciales problemas de rendimiento académico futuro. Por lo tanto, los niveles de consistencia hallados son aceptables para pruebas de grupo y de despistaje, que varían entre 0.60 y 0.80, respectivamente (Wasserman & Bracken, 2003). Considerando que la BDPG contiene características propias de una prueba de grupo y de despistaje, la magnitud de los coeficientes hallados mantiene un compromiso favorable para interpretar los puntajes asumiendo montos de error de medición tolerables.

Distribución de puntajes

Gradiente del ítem. La gradiente del ítem fue satisfactoria los puntajes subescalares del instrumento, excepto en el puntaje de Habilidades de Vocabulario. En este puntaje, los cambios entre las unidades de puntaje estandarizado fueron más numerosa que en los demás puntajes; los problemas de gradiente en este puntaje se concentraron más en las regiones extremas de la distribución. En cambio, para las demás subescalas la gradiente obtenida en los puntajes ha cumplido el criterio óptimo de gradiente del ítem, que es una diferencia igual o menor a 5 puntos entre cada punto obtenible. El amplio rango de puntajes que cumplen este criterio permite hacer diferenciaciones finas en el desempeño de los niños. Considerando que el BDPG se construyó para hacer interpretaciones nomotéticas, es decir, interpretaciones comparativas a un grupo de referencia (Weiner, 2003), los resultados muestran que en las secciones intermedias de los puntajes se pueden discernir más finamente la variabilidad del desempeño de un niño evaluado. Sin embargo, la diferenciación se hace menos precisa en los extremos de la distribución de los puntajes, especialmente en los puntajes debajo de dos desviaciones estándares de la media. La falta de una mayor discriminación en este nivel de puntuación se asoció, en primer lugar, al poco número de examinados con puntajes bajos en las muestras, haciendo que los cambios entre un punto y otro sean menos continuos. En segundo lugar, otro aspecto que influenció en la gradiente fue la presencia de puntajes extremos bajos, que generablemente se definen como outliers (Cohen, 2001) y que son valores únicos muy alejados de la densidad central de los datos. En tercer lugar, el efecto de la estimación de los puntajes más bajos fue que se obtuvieran puntajes estandarizados aproximados para un puntaje directo que no había sido logrado por el sujeto.

Las consecuencias de este problema de gradiente en la región extrema es que posiblemente ocurra una sobre-estimación de las capacidades de un niño, especialmente

cuando los niños obtengan muy bajos desempeños; es decir, los niños con puntajes de -1.5DE podrían parecer más hábiles de lo que en realidad son. Esto se debe a la inapropiada gradiente de los puntajes en esa zona de rendimiento, que produce cambios grandes entre uno y otro puntaje directo, y consecuentemente, entre cada unidad de puntaje estandarizado. Una alternativa razonable puede ser la inclusión de ítems más sencillos en la escala de puntuación debajo de 1.5 desviaciones estándares, el cual puede ser una estrategia que favorezca la interpretación normativa hacia los niños con menos capacidades aprendidas.

Los resultados del análisis de la gradiente del ítem utilizó la distribución de puntajes directos, el mismo que generalmente tienen menos continuidad entre cada punto en la escala de puntuación y cuya variación puede ocurrir por error de muestreo (Angoff, 1971; Thorndike, 1989). Una cuestión abierta no resuelta es la evaluación de la gradiente del ítem usando los puntajes ajustados a una distribución teórica, como la normal o beta-binomial de 4 parámetros. Ya que el ajuste de las distribuciones empíricas hacia estas distribuciones teóricas tiende a suavizar los cambios de la densidad de los puntajes, es plausible hallar gradientes apropiadas en los extremos de la distribución de puntajes. Esto es un punto de evaluación empírica que se resolverá en una futura investigación.

En general, los resultados, de este estudio señalan que los puntajes muestran adecuada sensibilidad en las distribuciones de los puntajes, para lograr una mejor diferenciación de las capacidades de los niños en casi todo el rango de puntajes; esta característica permitiría apoya la más exactitud.

Piso. El piso no ha sido estable en los años muestreados y en las subescalas, excepto Percepción Visual que ha mostrado un piso más estable. En el lado opuesto, Habilidad de Vocabulario presentó más problemas. Esto podría solucionarse con ítems más fáciles que puedan extender el escalamiento hacia los desempeños más bajos posibles

pero no iguales a un valor de cero. La inconstancia hallada en el piso de cada subescala puede representar una variación eventual de este aspecto debido a las características únicas de la muestra en cada periodo muestreado.

Una consecuencia de ello es que los puntajes con problemas en el piso pueden no ser sensibles a desempeños bajos, y obtener falsos negativos, ya que los niños de más bajo rendimiento presentarán puntajes más elevados debidos que no hay una mayor extensión hacia el extremo inferior de la distribución. Esto sugiere que los puntajes no podrían ser óptimos estimadores de la baja capacidad de un niño y producir una imagen inexacta del niño evaluado.

Esta dificultad, sin embargo, no podría ser una evidencia para concluir en que el BDPG tiene una limitación para su uso general. Esto es debido a los objetivos del BDPG; es decir, la finalidad del BDPG es dar información que sirva como un puente entre la dificultad del niño y la una exploración diagnóstica más completa. Como elemento de despistaje, por lo tanto, la información descriptiva que se obtendría sería una clasificación inicial del niño para su continuidad en su vida académica o su identificación como estudiante en riesgo de futuros problemas en el rendimiento académico. Esta capacidad de discriminación sería importante dentro de los propósitos de despistaje o screening, sin embargo, aunque una discriminación fina es importante y deseable, en el estado actual de la investigación del BDPG como herramienta de despistaje, sería suficiente determinar un punto de corte debajo de 1.5 desviaciones estándares para tomar decisiones de evaluación individual o derivar a tratamiento preventivo. Discriminaciones más finas pueden necesitarse en las evaluaciones individuales y más extensas desde las cuales se obtienen informaciones diagnósticas y prescriptivas más precisas. Aunque esta recomendación (punto de corte en -1.5DE) es racional más que empírica, se debe verificar en un estudio posterior.

De modo similar, se deja una cuestión abierta respecto al piso de las pruebas del BDPG, pues este puede ser evaluado luego del ajuste a una distribución teórica. Esto significa que los puntajes estandarizados pueden variar en los extremos para hacer que los cambios entre un punto y otro sea más "suave" (smoothing).

Finalmente, respecto al atributo de la gradiente del ítem se puede afirmar se ha observado una buena gradiente en los niveles medios de puntación en todas las escalas, y en que en varias ocasiones ha mostrado ser menos del criterio máximo recomendado. Esta situación favorece la discriminación fina del rendimiento en estas regiones de puntuación, que incluso se extiende algo más allá de una desviación estándar más allá de la media. Pero los problemas en la gradiente han ocurriendo constantemente en los extremos de la distribución de los puntajes, específicamente entre alrededor o más a una desviación estándar sobre la media, y luego alrededor o menos de dos desviaciones estándares debajo de la media. Pero en este aspecto psicométrico, el puntaje de Habilidades de Vocabulario ha tenido mayores distorsiones y ha sido más inestable en los años muestreados. Parece que una mayor posibilidad de obtener problemas en el apropiado diagnóstico preliminar y en descripción del niño ocurriría en los puntajes extremos, y esto es particularmente crítico para identificar a niños de muy bajos desempeño.

Por otro lado, el piso satisfactorio en los puntajes, excepto también en habilidades de Vocabulario, en que ha sido muy inestable entre los años muestreados. Pero en general, un piso lo suficientemente bajo asegura que se pueda aproximarse la estimación de niños bajo desempeño, y con seguridad requerirán apoyo psicopedagógico.

Propiedades normativas

Las propiedades normativas del BDPG fueron examinadas mediantes la exploración de las diferencias Una de las transformaciones comunes de los puntajes

empíricos de una distribución es que estos puntajes se transformen en una distribución normal, N(μ, σ^2), en que todos los valores de los puntajes podrían ser referidos a las probabilidades acumulativas de la curva normal para el cálculo de la ubicación de los sujetos evaluados (Howell, 1997). Muchas variables en la medición psicológica tienen esta forma distribucional o se aproximan a ellas (Anastasi & Urbina, 1997; Yang, 2007), y parece ser la distribución teórica más apropiada para hacer un ajuste de los datos empíricos. Sin embargo, un ajuste distribucional de los datos puede beneficiarse mejor si se incluye el monto de error de medición en los datos, y si teóricamente la variable empírica pertenece a una familia de distribuciones estadísticas más apropiadas. Por estos motivos, los puntajes fueron ajustados a un modelo beta binomial de cuatro parámetros (4PB). El efecto del ajuste para las subescalas fue eficiente, pues la bondad del ajuste no alcanzó significancia estadística luego de la corrección Bonferroni, excepto para el puntaje total. El puntaje total mostró un menor ajuste al modelo 4PB, y las diferencias entre la proporción empírica y la ajustada tendió a provenir de la distribución ubicada sobre la medida. En esta región, las diferencias entre las proporciones fueron algo más grandes que los puntajes debajo de la media. Este efecto podría contener también las diferencias de ajuste de las otras subescalas, pues el puntaje total consiste en la suma de estos puntajes, pero no se verificó esta afirmación. Este es un aspecto contrastable, así como la comparación de este ajuste con la distribución teórica normal y establecer el más eficiente modelo de distribución. Hay reportes previos que enfatizan que los modelos de distribución binomial, y mejor aún 4PB, son teórica y prácticamente más apropiados (Hanson & Brennan, 1990), pues el puntaje total se obtiene de la suma de respuestas dicotómicas, contiene información de la confiabilidad del puntaje en cuestión y son más flexibles comparado con el ajuste a una curva normal teórica.

Los puntajes de las subescalas y del puntaje total en general mostraron leves desviaciones de la normalidad, en un sentido asimétrico y curtósico. Esto parece una expresión de distribuciones inesperadas del rendimiento de los niños o de problemas en la construcción de los ítems. Sin embargo, asimetrías leves así como formas platicúrticas han sido características de la distribución de puntajes en pruebas de rendimiento académico (Cook, 1959; Lord, 1955), y parecen ser más comunes de lo que se podría esperar si se piensa que el rendimiento académico, así como otras variables asociadas al desarrollo cognitivo en población no clínica, se distribuye subyacentemente como la distribución teórica normal.

Al aplicar un procedimiento de obtención de normas, se tomó en cuenta las características distribucionales que tienden a ser más frecuentes en pruebas de rendimiento académico (Cook, 1959; Lord, 1955), y que justamente son distribuciones empíricas que no siguen una distribución teórica normal, situación que es más la norma que la excepción (Micceri, 1989). Aunque el ejercicio racional puede indicar que esta es quizás la mejor representación teórica de la distribución de puntajes, se halló que el ajuste al modelo 4PB fue satisfactorio para las subescalas, pero no para el puntaje total. La "maldad" de ajuste fue levemente observable considerando la razón entre el χ^2 y sus grados de libertad. Aun cuando las distribuciones de las subescalas parecen ser apropiadamente descritas por el modelo 4PB, el efecto acumulativo de las variaciones en los componentes del puntaje total puede ser la explicación plausible del desajuste. Esto efecto fue notorio considerando que el ajuste del puntaje total se hizo sobre el puntaje directo, derivado de la suma de las subescalas no ajustadas; por lo tanto, si el puntaje total previo al ajuste se hubiera realizado con los puntajes de las subescalas ajustados al modelo 4PB, es probable no se hubiera detectado el desajuste. Esto no fue evaluado en la presente investigación, y será un objetivo de futura contrastación.

La aplicación de este modelo, al ser teóricamente más apropiado, no es concurrente con la expectativa de hallar desviaciones positivas y negativas igualmente probables alrededor del valor central. Las desviaciones de la media no cayeron rápidamente en el grado en que se alejaban de la media, pues la tendencia platicúrtica de la distribución de los puntajes presentaba proporciones distribucionales ajustadas algo más elevadas que las proporciones originales, especialmente en la distribución de los puntajes sobre la media.

Uno de los efectos hallados cuando se efectuó el ajuste fue que la dispersión de los puntajes disminuyó ligeramente, y que los puntajes se hicieron más platicúrticos. Esta aparente contradicción se resuelve porque el modelo 4PB tiene como uno de sus parámetros la estimación de la confiabilidad, el mismo que se obtuvo del coeficiente KR-20 con modificación Horst (Horst, 1953; Charter, 1995). Este parámetro pudo haber afectado la estimación de las proporciones ajustadas en los extremos de la distribución, especialmente porque la restricción del rango de las distribuciones es más notoria en estas regiones debido a la disminución del tamaño muestral (Hurtz, 2008). Un efecto inmediato de la restricción del rango es la alteración de la consistencia interna (Feldt & Brennan, 1980), y por lo tanto, la precisión y la información de un puntaje se hacen menos fiables en estas regiones de la distribución.

Diferencias normativas

En general, se halló que los niños de los dos colegios muestreados provienen de una población en que las diferencias de rendimiento pueden ser consideradas triviales. Comparativamente, la interpretación normativa de ambos colegios puede prescindir del colegio de procedencia de los niños en las muestras del presente estudio. Como la variabilidad de los puntajes no provino de la pertenencia a determinados colegios,

entonces la aplicación de técnicas tradicionales de análisis (como la obtención de normas) en la muestra conjunta estaría garantizada (Peugh, 2009). Aunque esto no sugiere que las normas obtenidas pueden ser extensibles a otros colegios de niños en la misma edad y situación de recolección de datos, se podría suponer una razonable estabilidad normativa bajo condiciones similares respecto a la procedencia geográfica y cultural de los niños, y de distribución de horas y aprendizaje sea similar en los centros educativos de donde provienen. Las potenciales diferencias en los puntajes promedio en los colegios muestreados, y en general en otros colegios, puede provenir de varias condiciones intrínsecas y extrínsecas al niño, y de la interacción entre ellas. Intrínsecamente, las capacidades cognitivas del niño, así como la integridad neuronal y condiciones vulnerables pueden influenciar disminuir legítimamente en el rendimiento y aprendizaje de las habilidades pre-académicas, y no ser explicadas por error de medición; por otro lado, las condiciones ambientales físicas y culturales de su familias, la capacidad económica y la calidad de la atención emocional o orientadas al rendimiento por parte de los apoderados, todo ello aporta una varianza explicativa relevante a las potenciales diferencias. Obviamente, la interacción entre estas condiciones tendrá un efecto en la velocidad y alcance de los aprendizajes de los niños, y distribucionalmente los niños con estas vulnerabilidades extrínsecas e intrínsecas podrían representar menos de 1 desviación estándar debajo de la media.

Hallar diferencias de grupo tiene una consecuencia importante en la toma de decisiones para intervención preventiva, pues por ejemplo, si el rendimiento de los niños ingresantes al primer grado de un colegio se diferencia del grupo normativo en cerca o más de 0.8 unidades estándares (en términos de magnitud del efecto, d Cohen) debajo de la media, la intervención es una exigencia consecuente de esta evaluación. Ya que los puntajes interpretados provienen de procedimientos de escalamiento en que se incluye la

precisión de los puntajes (usando la confiabilidad en el modelo beta binomial de 4 parámetros [4PB]), la identificación de los niños en riesgo de problemas da un respaldo para que las decisiones de intervención sean generalmente confiables. La recuperación en las áreas deficitarias, y el fortalecimiento de las capacidades en el grupo de riesgo serían los objetivos de intervención.

Además de la oportunidad que tiene el niño de ser intervenido oportunamente, la apropiada información obtenida de los puntajes del BDPG da un respaldo a su validez consecuencial, pues favorece el proceso de tomar decisiones en un periodo donde el niño tiene mejor oportunidad de recuperar los aprendizajes no logrados (VanDerHeyden et al., 2001), e influenciar en su rendimiento académico.

Aunque el BDPG puede ser utilizado para apoyar las decisiones de intervención, el objetivo principal es la identificación temprana en una metodología de evaluaciones de despistaje (screening). Por lo tanto, su mejor y más apropiado uso será la derivación de los niños detectados hacia una evaluación más minuciosa y de carácter diagnóstico. Pero en la situación de no disponer los medios suficientes para una evaluación individual de este tipo, los puntajes del BDPG pueden ser orientativos para intervenir rápidamente o diferenciar grupos de niños en riesgo de bajo rendimiento en lectura y matemáticas. Por ejemplo, en los primeros años dos grados de primaria, se sugiere que identificación de niños disléxicos tome en cuenta la diferenciación con aquellos niños de lento inicio en la lectura (Badian, 1993). Los niños de lento inicio desarrollan habilidades relativamente normales en los años posteriores de escolaridad. Y son niños que cumplen con los criterios de discapacidad lectora en varias componentes de la lectura pero que llegan posteriormente a ser lectores normales; generalmente están asociados con pobre alfabetización en el hogar, baja habilidad lectora en las madres, y bajo nivel socioeconómico.

Las comparaciones normativas en ambos colegios arrojaron similaridades en los puntajes promedio, tanto en las comparaciones intra-año como en las comparaciones entre-años. Por lo tanto, estos dos niveles de análisis revelaron que la varianza de los puntajes se debió a las diferencias en las unidades básicas de análisis (los niños). Excepto las diferencias halladas en tres de las subescalas en uno de los periodos de evaluación (año 2004), la variabilidad en el resto de periodos puede observarse como "ruido" (Peugh, 2009). Por este motivo, la obtención de la distribución de normas ajustadas modeladas por la distribución 4PB no se consideró apropiada para las estas escalas, y en general, los datos de este periodo no fueron ingresados para aproximarse a las normas del BDPG.

Implicaciones para la práctica

Las evidencias favorables para el instrumento validado en el presente estudio dan respaldo para su uso planificado desde su construcción: evaluación de despistaje de habilidades pre-académicas en niños que ingresan al primer grado de primaria. Por lo tanto, puede ser valorado como una herramienta de elección frente a los existentes, los que generalmente son antiguos, tienen pobre respaldo psicométrico y no garantizan resultados confiables ni válidos. Por otro lado, el instrumento favorece un balance costo-beneficio sobre la rapidez de su aplicación y calificación, la objetividad de los resultados, el respaldo científico de su construcción, y las consecuencias de su uso.

Aunque el BDPG permite la sencillez de su aplicación en grupos de los niños, se debe tener especial cuidado en los puntajes bajos, dado que la presente investigación demuestra que pueden ser menos precisos. Este problema ha sido previsto en algunos aportes descriptivos sobre pruebas psicológicas en el área educativa (Gregory, 2000). Sumado a lo anterior, las exactas instrucciones propuestas en el presente trabajo abre el camino para su aplicación estandarizada, guiando al usuario para que realice las mínimas modificaciones que puedan ser fuente de error y alteración en los puntajes, las cuales muchas veces pasan inadvertidos (Lee et al., 2003).

Dado que las pruebas cognitivas en las edades tardías del periodo preescolar (4-6 a 5-6) tienden a ser psicométricamente más sólidas (Alfonso & Flanagan, 1999), el instrumento adaptado en el presente estudio posee la garantía adicional de mostrar propiedades aceptablemente invariantes. Esto fue probado parcialmente en las comparaciones entre los años muestreados, y por lo tanto ofrece una perspectiva optimista sobre su continuo uso.

Limitaciones

Las potenciales limitaciones del estudio es la representatividad de la muestra. Aunque se asumió similaridad demográfica (que fue corroborada empíricamente), la se puede afirmar de la similaridad psicológica. Esta situación limita la generalizabilidad de nuestros resultados, y su poder descriptivo hacia instituciones educativas en otros distritos. El límite, en este punto, también se relaciona con la ubicación general y específica del grupo de niños muestreados, ya que todos viven en Lima. Pero el tamaño de la muestra, y la comprobación de la poco variable características demográficas pueden sugerir una potencial extensión de nuestros resultados.

La integración con estrategias cualitativas no se ha efectuado en este estudio, ni como estrategia para la validez de contenido. Dado que el contenido del instrumento adaptado es la fuente esencial para los posteriores análisis, lo obtenido del análisis de contenido elegido es solo una fuente parcial e incompleta de este aspecto de la validez.

Recomendaciones

a) La administración grupal de pruebas de despistaje, como la BDPG, debe ser cuidadosamente escrutinizada, ya que tiene un efecto directo sobre la calidad del servicio de evaluación, el error de medición y finalmente en la estimación del rendimiento del niño. Se recomienda la elaboración de un protocolo que defina la secuencia de administración, los eventos presentados, y el control de eventos aleatorios dentro de un plan de evaluación.

b) Las medidas de habilidades convergentes y divergentes con el BDPG han sido principalmente de despistaje o parcialmente cubrían el dominio de contenido de sus constructos. Elegir medidas más complejas y con mayor información diagnóstica daría un mejor soporte al BDPG para estimar atributos que sean útiles para describir el desempeño actual. Por ejemplo, medidas de percepción visual con un mayor muestreo de contenido que el usado en nuestro estudio (Test de Percepción Visual No Motor), demostrarían un indicador de validez para el uso prescriptivo del BDPG.

c) El uso de instrumentos como el BDPG no debería aislarse del beneficio directo en los niños detectados con potenciales problemas de rendimiento. De este modo, la utilidad del instrumento es consecuente al impacto en la evaluación más precisa, y en ausencia de esta, en la rehabilitación de los déficits posiblemente detectados. Por lo tanto, aún con posterior investigación psicométrica con el BDPG, se debe incluir estrategias de evaluación adicional con poco costo, y de intervención preventiva.

d) Una pregunta que queda abierta es la del impacto del contexto cultural y familiar del niño sobre sus puntajes en el BDPG. Aunque son variables distales, pueden

constituir un conjunto heterogéneo de indicadores que explican una parte de la varianza en el BDPG. Ya que no se ha probado ni directa ni indirectamente esta probable influencia en esta investigación, se espera que el diseño de investigación posterior incluya esta variable.

e) Las evidencias predictivas del BDPG han sido satisfactorias, pero el periodo en que ocurrió el criterio de rendimiento ha sido breve respecto a la aplicación del BDPG, y debería del niño puede extenderse más. De tal manera, la capacidad predictiva de las habilidades medidas en el BDPG pueden probar ser también predictoras del rendimiento del niño. Así que, se recomienda evaluar longitudinalmente hasta el tercer grado, el desempeño de los niños identificados con potenciales problemas de rendimiento.

f) Ya que la recolección de datos se hizo en un contexto profesional de evaluación de niños, los resultados presentados pueden haber servido para tomar decisiones y modificar la percepción que las profesoras tienen sobre sus los niños evaluados. Por lo tanto, se recomienda evaluar el impacto de la evaluación y sus resultados en las prácticas instruccionales, así como en la percepción de la profesora sobre el comportamiento académico del niño.

g) La interpretación de los puntajes bajos deben ser considerados con cuidado, especialmente en la aplicación grupal del BDPG.

REFERENCIAS

Abarca, S. et al. (1965). *Adaptación y análisis estadístico del Metropolitan Readiness Test de Hildreth G. y N. Griffith*. Memoria para optar al título de Psicólogo. Santiago de Chile: Universidad Católica de Chile.

Abramson, J. H. (2004). WINPEPI (PEPI-for-Windows) computer programs for epidemiologists. *Epidemiologic Perspectives & Innovations, 1(*6), Disponible en línea: www.epi-perspectives.com/content/1/1/6.

Albiragorta, F. J. (1983). *Estudio comparativo y correlativo de la madurez para la lecto-escritura* y el ajuste emocional en niños de 1er. grado. Tesis para optar el Título de Psicología, Facultad De Psicología. Universidad Inca Garcilaso de la Vega. Lima: UIGV.

Alf, E. F., Jr., & Abrahams, N. M. (1975). The use of extreme groups in assessing relationships. *Psychometrika, 40*, 563-572.

Alfonso, V. C., & Flanagan, D. P. (1999). Assessment of cognitive functioning in preeschoolers. In E. V. Nutall, I. Romero, & J. Kalesnik (Eds.), *Assessing and screening preschoolers* (2nd. ed., pp. 186-217). New York: Allyn & Bacon.

Alfonso, V. C., & Flanagan, D. P. (2006). Best practices in the use of the Standford-Binet Intelligence Scales, Fifth Edition (SB5) with preschoolers. In B. A. Bracken & R. Nagle (Eds.), *Psychoeducational Assessment of preschool children* (4th ed., pp. 267-295). Mahwah, NJ: Erlbaum.

Alzamora, C. (1981*). Nivel socioeconómico en el desarrollo psicomotriz de niños de Lima Metropolitana.* Tesis para optar el Título de Psicología, Facultad De Psicología. Lima: USMP.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (1999). *Standards for educational*

and psychological testing. Washington, D.C.: American Psychological Association.

Anastasi, A., & Urbina, S. (1997). *Psychological testing (7th ed)*. New York: McMillian.

Angoff, W. H. *(*1971*).* Scales, norms, and equivalent scores. In R. L. Thorndike. *(*Ed.*),* *Educational measurement* (*2nd ed*.), pp. 508-600*).* Washington, DC*:* American *Council on* Education.

Aranda, V. (1974). E*standarización de la prueba de madurez mental de california, serie pre-primaria*. Tesis para optar la titulación en Psicología, Facultad de Psicología. Universidad Nacional Mayor de San Marcos. Lima: UNMSM.

Ardila, R. (2004). Psicología Latinoamericana: El primer medio siglo. *Revista Interamericana de Psicología, 38*(2), 317-322.

Arraya N., M. (1992). *Estudio descriptivo-comparativo de la coordinación visomotora en niños de 5 a 7 años del nivel inicial con problemas de aprendizaje*. Tesis para optar el Título de Psicología, Facultad De Psicología. Lima: USMP.

Badian, N. A. (1993). Prediction reading progress in children receiving special help. *Annals of Dyslexia, 43,* 90-109.

Badian, N. A. (1994). Preschool Prediciton: Orthographic and phonological skills, and reading. *Annals of Dyslexia, 44,* 3-24.

Badian, N. A. (2001). Phonological and orthographic processing: Their roles in reading prediction. *Annals of Dislexia, 51,* 179-202.

Basilio, C., Puccini, R., Silva, E., & Pedromonico, M. (2005). Living conditions and receptive vocabulary of children aged two to five years. *Revista de Saúde Pública, 39*(5), 725-730

Beery, K. E. (2000). *Prueba Beery-Buktenica del desarrollo de la Integración Visomotriz (VMI) (4ta. ed.)*. México, D. F: El Manual Moderno.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*(2), 238-246.

Bentler, R. M., & Wu, E. J. C. (2004). *EQS 6.1 for windows* [Statistical Program]. Encino, CA: Multivariate Software, Inc.

Berdicewski, O. & Milicic, N. (1978). *Prueba de Funciones Básicas*. Santiago: Galdoc.

Berdicewski, O., & Milicic, N. (1974). Jardín infantil y su influencia en el rendimiento de coordinación visomotora, discriminación auditiva y lenguaje, medidos con una prueba de funciones básicas. *Revista Chilena de Pediatría, 45*(6), 505-508.

Berdicewski, O., & Milicic, N. (1974). Una prueba para detectar a los niños con alto riesgo de presentar problemas en el aprendizaje en primer año básico. *Revista Chilena de Pediatría, 45*(6), 513-515.

Berdicewski, O., & Milicic, N. (2004). *Prueba de Funciones Básicas* (35ta. ed.). Santiago: Editorial Universitaria.

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Spurious evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105*, 467-477.

Boehm, A. E. (2000). *Boehm Test of Basic Concepts, Third Edition.* San Antonio, TX: Psychological Corporation

Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. *Journal of Psychoeducational Assessment, 5,* 313-326

Bracken, B. A. (1998). *Examiner's manual for the Bracken Basic Concept Scale–Revised.* San Antonio, TX: Psychological Corporation.

Bracken, B. A. (2000). Maximizing construct relevant assessment: The optimal preschool testing situation. In B. A. Bracken (Ed.), *The psychoeducational assessment of preschool children* (3rd ed., pp. 33-44). Needham Heights, MA: Allyn & Bacon.

Bracken, B. A. (2002). *Bracken School Readiness Assessment (BSRA)*. Texas: The Psychological Corporation.

Braden, J. P. (1986). Testing correlations between similar measures in small samples. *Educational and Psychological Measurement, 46*, 143-148.

Bradley-Johnson, S., & Durmusoglu, G. (2005). Evaluation of floors and item gradients for reading and math tests for young children. *Journal of Psychoeducational Assessment, 23*, 262-278.

Brannigan, G. G., & Brunner, N. A. (2002). *Guide to the Qualitative Scoring System for the modified version of the Bender-Gestalt Test (2nd ed.)*. IL: Charles C. Thomas.

Brannigan, G. G., & Decker, S. L. (2003). *Bender-Gestalt II Examiner's manual*.. Itasca, IL.: Riverside Publishing.

Bravo, L. (1997). Prueba experimental pre-lectora (PPL). *Boletín de Investigación Educacional, 12***,** 79-90.

Bravo, L. (2002). La conciencia fonológica como una zona de desarrollo próximo para el aprendizaje inicial de la lectura. *Estudios Pedagógicos, 28,* 167-177.

Bravo, L. (2004). La conciencia fonológica como una posible "zona de desarrollo próximo" para el aprendizaje de la lectura inicial. *Revista Latinoamericana de Psicología, 36*(1), 21-32.

Brennan, R. L. (2004). *Manual for BB-CLASS: A computer program that uses the betabinomial model for classification consistency and accuracy, Version 1.1* (CASMA Research Report No. 9). Iowa City: University of Iowa.

Brenner, A. (1967). *Anton Brenner Developmental Gestalt Test of School Readiness*. Los Angeles, CA: Western Psychological Services.

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the develoment and evaluation of personality scales. *Journal of Personality, 54*, 106-148.

Brown, L., Sherbenou, R. J., & Johnsen, S. K. (2000). *Test de Inteligencia No Verbal (TONI-2)* (2da. ed.). Madrid: TEA.

Byrne, B. (2010). *Structural Equation Modeling With AMOS: Basic Concepts, Applications, and Programming (Multivariate Applications), 2nd edition*. New York: Taylor & Francis.

Cadieux, A., & Boudreault, P. (2002). Psychometric properties of a kindergarten behavior rating scale to predict later academic achievement. *Psychological Reports, 90*, 687-698.

Calderón, J. (1991). *Relación entre el examen de ingreso a Primer Grado y el rendimiento escolar en un grupo de niños de Primer Grado de Educación Básica Regular de un colegio particular de Lima Metropolitana*. Tesis para optar el Título de Psicología, Facultad De Psicología. Lima: UNIFE.

Canales, R. (1990). *Estudio exploratorio con la Prueba de Pre-cáculo Milicic y Schmidt (1982) en población de clase media de la zona urbana de Huancayo*. Tesis para optar el Título de Psicología, Facultad De Psicología. Lima: UNIFE.

Canivez, G. L., Willenborg, E., & Kearney, A. (2006). Replication of the Learning Behaviors Scale factor structure with an independent sample. *Journal of Psychoeducational Assessment, 24*, 97-111.

Caskey, W. R., Jr., & Larson, G. L. (1975). Two modes of administration of the Bender Visual-Motor Gestalt Test to kindergarten children. *Perceptual and Motor Skills, 45*(1), 1003-1006.

Chan, P. W. (2000). Comparison of visual motor development in Hong Kong and USA assessed on the Qualitative Scoring System for the Modified Bender Gestalt Test. *Psychology Reports, 88,* 236-240.

Chan, P. W. (2002). Relationship of the visual motor development and academic performance in young children in Hong Kong assessed in the Bender-Gestalt Test. *Perceptual and Motor Skills, 90,* 209-214.

Charter, R. A. (1995). The under-representation of Horst's modification of the KR-20 reliability coefficient. *Perceptual and Motor Skills, 81,* 770.

Charter, R. A. (1997). Effect of measurement error on test of statistical significance. *Journal of Clinical and Experimental Neuropsychology*, *19*(3), 458-462.

Charter, R. A. (2003). A breakdown of reliability coefficients by test type and reliability method, and clinical implications of low reliability. *The Journal of General Psychology, 130*(3), 290-304.

Chen, F. F., Sousa, K. H., & West. S. G. (2005). Testing measurement invariance of second-order factor models. *Structural Equation Modeling, 12,* 471-492.

Chen, Y., & Popovich, P. M. (2002). *Correlation: Parametric and Nonparametric measures* (Sage University). Thousand Oaks, CA: Sage.

Chumbipuma, M. (1996). *Madurez social y madurez para el aprendizaje de la lecto-escritura en niños del distrito de Santa Anita – Lima*. Tesis para optar el Título en Psicología, Facultad de Psicología. Lima: UNFV.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284-290.

Cicchetti, D.V., & Sparrow, S.S. (1981). Developing criteria for establishing the interrater reliability of specific items in a given inventory. *American Journal of Mental Deficiency, 86*, 127-137.

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in scale development. *Psychological Assessment, 7*(3), 309-319.

Cohen, B. H. (2001). *Explaining psychological statistics* (2nd. ed.). New York: John Wiley & Sons.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159.

Collarusso, R. P., & Hammill, D. D. (1980). *Test de Percepción Visual No Motriz: TPVNM*. Buenos Aires: Panamericana.

Condemarín, M. (1989). *Lectura Temprana: Jardín y primer grado*. Santiago: Andrés Bello.

Condemarín, M., Chadwick, M., & Milicic, N. (1981). *Madurez Escolar*. Santiago: Andrés Bello.

Cone, J. D. (1999). Observational assessment: Measure development and research issues. In P. C. Kendall, J. N. Burcher, & G. N. Holmbeck, *Handbook of Research Methods in Clinical Psychology (2nd ed.)*, (pp. 183 - 223). NY: John Wiley & Sons.

Cook, D. L. (1959). A replication of Lord's study of skewness and kurtosis of observed test-score distributions. *Educational and Psychological Measurement, 19*, 81-87.

Cope, R. T., & Kolen, M. J. (1990). *A study of methods for estimating distributions of test scores*. ACT Research Report 90-5. Iowa City, IA: ACT, Inc.

Correa, J. (1977). Madurez para el aprendizaje, aspectos teóricos y prácticos: Aplicación de la Prueba Jordan y Massey en niños de P. J. de Chorrillos. Tesis de Bachiller en Psicología, Facultad de Psicología.. Lima: UNFV.

Cronbach, L. J., Schönemann, P., & McKie, D. (1965). Alpha coefficients for stratified parallel tests. *Educational and Psychological Measurement, 25*, 291-312.

Curran, P. J., West, S. G, & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods, 1*, 16-29.

D'Agostino, R.B., & Cureton, E.E. (1975). The 27 percent rule revisited. *Educational and Psychological Measurement, 35*, 47-50.

Dávila, M., Tello, K., & Merino, C. (2009). Nueva versión del Dibujo de una Persona para la Estimación Intelectual: Estudio psicométrico preliminar en preescolares. *Fractal: Revista de Psicología, 21*(2), 443-444.

De Bruin, G. P. (2004). Problems with the factor analysis of items: solutions based on item response theory and item parceling. *South African Journal of Industrial Psychology, 30*(4), 16-26.

Doig, B. (2005). Developing formal mathematical assessment for 4- to 8-year-olds. *Mathematics Education Research Journal, 16*(3), 10-119.

Dumont, R., Willis, J. O., (2007). Bracken Basic Concept Scale-Revised. In C. R. Reynolds, & E. Fletcher-Janzen (Eds.), *Encyclopedia of special education: A reference for the education of children, adolescents, and adults with disabilities and other exceptional individuals* (3th ed., pp. 316-317). Hoboken, NJ: John Wiley and Sons.

Dumont, R., Willis, J. O., (2007). Bracken Basic Concept Scale-Revised. In C. R. Reynolds, & E. Fletcher-Janzen (Eds.), *Encyclopedia of special education: A reference for the education of children, adolescents, and adults with disabilities and other exceptional individuals* (3th ed., pp. 316-317). Hoboken, NJ: John Wiley and Sons.

Educational Testing Service. (1989). *Reading readiness. Annoted bilbliography of test*. Princeton, NJ: Educational Testing Service.

Egeland, B. (1970). School Readiness Survey by F. L. Jordan, James Massey (Review). *Journal of Educational Measurement, 7*(1), 58-59.

Emmons, M. R., & Alfonso, V. C. (2005). A critical review of the technical characteristics of current preschool screening batteries. *Journal of Psychoeducational Assessment, 23*(2), 111-127.

Erford, B. T. (2004). The Reading Essential Skill Screener - Preschool Version (RESS-P): Studies of reliability and validity. *Assessment for Effective Intervention, 29*(3), 19-34.

Erford, B. T., & Biddison, A. R. (2006). The Math Essential Skills Screener - Upper Elementary Version (MESS-U): Studies of reliability and validity. *Assessment for Effective Intervention, 31*(4), 33-46.

Espinoza, J., & Matos, T. (1991). *Los procesos perceptivos visuales en niños de educación inicial 1er. y 2do. Grado de Primaria de colegios estatales del distrito de La Molina*. Tesis para optar el Título en Psicología, Facultad de Psicología. Lima: UNIFE.

Espinoza, J., Piedra, M., & Sotomarino, J. (1995). *Estandarización de la Prueba de Funciones Básicas para la lectura y escritura en Lima Metropolitana*. Tesis para optar el Título en Psicología, Facultad de Psicología. Lima: UNIFE.

Eyzaguirre, N. (1996). *Evaluación de Habilidades para el Aprendizaje: Manual de 1ro al 3er grado*. Lima. CEDAPP/Rädda Barnen.

Fan, X., B. Thompson, & L. Wang (1999). Effects of sample size, estimation method, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling, 6*, 56-83.

Feldt, L. S. (1961). The use of extreme groups to test for the presence of a relationship. *Psychometrika, 26*, 307-316.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. En: R. L. Linn (Ed.), *Educational Measurement* (3rd ed.) (pp. 105-146). New York: American Council on Education and MacMillan.

Feldt, L. S., & Charter, R. A. (2003). Estimating the reliability of a test split into two parts of equal or unequal length. *Psychological Methods, 8*(1), 102-109.

Feldt, L. S., Woodruff, D. J. & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement, 11*, 93-103.

Filho, L. (1960). *Test ABC de verificación de la madurez necesaria para el aprendizaje de la lectura y escritura (6a. Ed.)*. Buenos Aires: Kapelusz.

Flanagan, D. P., & Alfonso, A. C. (1995). A critical review of the technical characteristics of new and recently revised intelligence tests for preschool children. *Journal of Psychoeducational Assessment, 13,* 66-90.

Flanagan, D. P., & Alfonso, V. C. (2005). A critical review of the technical characteristics of current preschool screening batteries. *Journal of Psychoeducational Assessment, 23*(2), 111-127.

FONCODES (2006). *Focalización geográfica: Nuevo mapa de pobreza de FONCODES 2006*. Lima: Ministerio de la Mujer y Desarrollo Social

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50.

Fowler, M. G. & Cross, A. W. (1986) Preschool risk factors as predictors of school performance. *Journal of Developmental and Behavioral Pediatrics, 7*, 237–241.

Fowler, M.G. & Cross, A.W. (1986) Preschool Risk Factors as Predictors of School Performance. *Journal of Developmental and Behavioral Pediatrics, 7*, 237–241.

Fowler, R. L. (1992). Using the extreme groups strategy when measures are not normally distributed. *Educational and Psychological Measurement, 16*(3), 249-259.

Frostig, M. (1980). *Método de evaluación de la percepción visual*. México D.F.: El Manual Moderno.

Fuller, G. B., & Vance, B. (1995). Interscorer reliability of the modified version of the Bender-Gestalt Test for Preschool and Primary School Children. *Psychology in the Schools, 32*(4), 264-266.

Furlan, L. & Alderete, A. (2004). Diagnóstico de habilidades básicas para el ingreso a primer grado en niños de zonas urbano-marginales y rurales. *Evaluar, 4*, 70-94.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction* (6ta ed.). White Plains, NY: Longman Publishers.

Gamache, L. M. (1987). *A Procedure to adjust individual scores for construct invalidity*. ACT Research Report Series, Nro. 87-18.

Gardner, M. F. (2000) *Cognitive (Intelligence) Test: nonverbal (CIT:nv)*. Hydesville, CA: Psychological and Educational Publications.

Garg, R. (1985). An empirical comparison of three strategies used in extreme group designs. *Educational and Psychological Measurement, 43*, 359-371.

Garret, H. B. (1971). *Estadística en Psicología y Educación*. Buenos Aires: Paidós.

Gastelumendi, E., Isasmendi, A., Slovak, G., & Semeleng, Z. (1977). *Test 5-6, Forma B.* Montevideo: Kapeluz.

Gilmer, J. S., & Feldt, L. S. (1983). Reliability estimation for a test with parts of unknown lenghts. *Psychometrika, 48*(1), 99-111.

Gracey, C. A., Azzara, C. V., & Reinherz, H. (1984). Screening revisited: A survey of U.S. requirements. *Journal of Special Education, 18*(2), 101-107.

Graciela, C. (2006). *Saberes Previos para el ingreso al 1er grado*. Documento no publicado. I. E. María Inmaculada. Chorrillos, Lima: Autor.

Gregory, R. J. (2000). *Psychological testing: History, principles, and applications* (3rd ed.) Boston: Allyn and Bacon.

Grigorenko, E. L. & Sternberg, R. J. (1999). *Assessing cognitive development in early childhood. Education: Early child development*. Washington, DC: World Bank.

Guevara, Y., & Macotela, S. (2000). Proceso de adquisición de habilidades académicas: una evaluación referida a criterio. *Revista Iberpsicología, 5*(2). Disponible en http://fs-morente.filos.ucm.es/publicaciones/iberpsicologia/Iberpsico9/guevara/guevara.htm.

Guevara, Y., Hermosillo, A., García, G., Delgado, U., & López, A. (2007). Evaluación y análisis del desarrollo académico en alumnos de primer grado de primaria, en H. Hickman y O. Tena (Ed.). *Jornadas del Proyecto de Aprendizaje Humano.* Universidad Nacional Autónoma de México

Gutierrez, N. C. (1988). *Exploración del nivel de madurez para el aprendizaje de la lectura en un grupo de niños de primer grado de Educación Básica Regular*. Tesis para optar el Título en Psicología, Facultad de Psicología. Lima: UPIGV.

Haeussler, I. M., & Marchant, T. (2003) *Test de Desarrollo Psicomotor 2-5 Años (10ma ed.)*. Santiago: Universidad Católica.

Hall, R. J., Snell, A. F., & Foust, M. S. (1999). Item parceling strategies in SEM: Investigating the subtle effects of unmodeled secondary constructs. Organizational Research Methods, 2(3), 233-256.

Hanson, B. A. (1991). *Method of moments estimates for the four-parameter beta compound binomial model and the calculation of classification consistency indexes.* ACT Research Report 91-5. Iowa City, IA: ACT, Inc.

Hanson, B. A., & Brennan, R. L. (1991). An investigation of classification consistency indexes estimated under alternative strong true scores models. *Journal of Educational Measurement, 27*(4), 345-359.

Harris, D. B. (1963). *Children's drawings as measures of intellectual maturity: A revision and extension of the Goodenough Draw-A-Man Test*. New York: Harcourt Brace Jovanovich.

Hasbrouck, J. (1990). Preschool assessment. In G. Tindal & D. Marston (Eds.), *Classroom-based Assessment: Testing for Teachers* (pp. 273-291). Columbus, OH: Merrill.

Hattie, J.A. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139-164.

Hildreth, G. H., Griffiths, N. L., & McGauran, M. (1965). *Metropolitan Readiness Tests*. New York: Harcourt, Brace and World, Inc.

Hintze, J. (2007). *NCSS, PASS, and GESS*. Kaysville, Utah: NCSS.

Hirsh-Pasek, K., Hyson, M. C., & Rescorla, L. (1990). Academic environments in preschool: Do they pressure or challenge young children. *Early Education and Development, 1*(6), 401-423.

Hopkins, B. (2005). Neuromaturational theories. In: Hopkins, B. (Ed.), The Cambridge encyclopedia of child development (pp. 37-48). Cambridge: Cambridge University Press.

Hopkins, K. D. & Weeks, D. L. (1990). Tests for normality and measures of skewness and kurtosis: their place in research reporting. *Educational and Psychological Measurement*, 50, 717-729.

Horst, P. (1953). Correcting the Kuder-Richardson reliability formula for dispersion of item difficulties. *Psychological Bulletin, 50*, 371-374.

Howell, D. C. (1997). *Statistical methods for psychology* (4th. ed.). Belmont, CA: International Thomson Publishing.

Huber, B. (1999, March). *Fitting distributions to data: Practical issues in the use of probabilistic risk assessment*. Paper presented in Quantitative Decisions, Sarasota, FL.

Hurtz, G. (2008, Aphril). Clarification on some misconceptions about conditional standard errors of measurement. In Hurtz, G. M. (Chair), *Integrating conditional standard errors of measurement into personnel selection practices*. Symposium presented at the Annual Conference of the Society for Industrial and Organizational Psychology, San Francisco, CA, April.

Ilg, F. L., Ames, L. B., Haines, J., & Gillespie, C. (1978). *Tests de madurez escolar Instituto Gesell*. Buenos Aires: Paidos.

Ilg, F. L., Ames, L. B., Haines, T., & Gillespie, C. (1978). *School readiness: Behavior tests used at the Gesell Institute* (Rev. ed.). New York: Harper & Row.

International Test Commission (2001). International guidelines for test use. *International Journal of Testing, 1*, 93-114.

Jackson, W. (1999). *Methods doing social research* (2nd. ed.). Scarborough, Canada: Prentice-Hall.

Jordan, F. L. & Massey, j. (1967). *School Readiness Survey (2nd ed.)*. Palo Alto, CA: Consulting Psychologists Press.

Keats, J. A. (1957). Estimation of error variances of test scores. *Psychometrika, 22*, 29-41.

Kenny, D. & Izard, J. (1993, Julio). *The development of a teacher rating scale for the early identification of children at risk of educational failure*. Ponencia presentada en el 23er Congreso internacional de Psicología Aplicada, Madrid, España.

Kenny, D. T. (1992). Can teachers be tests? A comparison of teacher ratings and test assessments of early reading performance. In H. Motoaki, J. Misumi, & J. B. Wilport (Eds.). *Social, educational and clinical psychology* (Vol. 3, pp. 177-178). London: Lawrence Erlbaum Associates.

Kenny, D. T., & Chekaluk, E. (1993). Early reading performance: A comparison of teacherd-based and test-based assessments. *Journal of Learning Disabilites, 26*(4), 227-236.

Kolen, M. J. (1991). Smoothing methods for estimating test score distributions. *Journal of Educational Measurement*, 28 (3), 257-282.

Köppitz, E. M. (1984). *El test gestáltico visomotor para niños* (10ma. ed.). Bs. As: Guadalupe.

Krasa, N. (2007). Is the Woodcock-Johnson III a test for all seasons?: Ceiling and Item Gradient Considerations in Its Use With Older Students. *Journal of Psychoeducational Assessment, 25*(1), 3-16.

Kulp, M. T. (1999). Relationship between visual motor integration skill and academic performance in kindergarten through thid grade. *Optometry and Vision Science, 73*(3), 159-163.

Kulp, M. T., & Schmidt, P. P. (1996). Visual predictors of reading performance in kindergarten and first grade children. *Optometry and Visual Science, 73*, 255-262.

Lee, D., Reynolds, C. R., & Willson, V. (2003). Standardized test administration: Why bother? *Journal of Forensic Neuropsychology, 3*(3), 55-81.

Lee, W. C., Hanson, B. A., & Brennan, R. L. (2002). Estimating consistency and accuracy indices for multiple classifications. *Applied Psychological Measurement, 26*(4), 142-432.

Levy, K. J. (1977). Non-normality and testing that a correlation equals zero. *Educational and Psychological Measurement, 37*, 691-694.

Lin, L. I. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics, 45*, 255-268.

Lindley, P., Bartram, D., & Kennedy, N. (2005). *EFPA review for the description and evaluation of psychological test: Test review form and notes for reviewers (version 3.41)*. Unpublished report.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: exploring the question, weighing the merits. *Structural Equation Modeling, 9*(2), 151–173.

Llactahuacchaca, G. (1994). *Comparación entre el desarrollo de la percepción visual en niños de la zona urbana y zona rural que ingresan al 1er grado en el distrito de Santa Cruz – Cajamarca*. Tesis para optar el Título en Psicología, Facultad de Psicología. Lima: UNFV.

López, J. (1995). *Teoría de la respuesta al ítem: Fundamentos*. Barcelona: Promociones y Publicaciones Universitarias.

Lord, F. M. (1955a). A survey of observed test-score distributions with respect to skewness and kurtosis. *Educational and Psychological Measurement, 15*, 383-389.

Lord, F. M. (1955b). Estimating test reliability. *Educational and Psychological Measurement, 15*, 325-336.

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika, 30*, 239–270.

Majluf, A. (1984). Rendimiento intelectual de niños – infantes a adolescentes de clases socioeconómicas media y baja de lima y de algunas provincias. *Revista de Psicología (PUCP), 2*(1-2), 57-73.

Majluf, A., Gutierrez, R., Delgado, I., & Torres, L. (1971). *Maduración para el aprendizaje de los niños de cinco años de las zonas marginales*. Anales del 2do Congreso de Psiquiatría, Asociación Psiquiátrica Peruana, Noviembre. Lima.

Mann, V. A., & Liberman, I. Y. (1984). Phonological Aawareness and verbal short-term memory. *Journal of Learning Disabilities, 17*(10), 592-599.

Manterola, A., Avendaño, P., Cotroneo, J., Avendaño, A., Valenzuela, C. (1986) Factores de riesgo en las dificultades de aprendizaje escolar en niños de medio económico-social medio y bajo. *Revista Chilena de Pediatría, 57*(4), 318-24.

Marchena, C. & Santos, M. (1986). Aprestamiento, madurez y lecto-escritura en niños de un Centro Educativo urbano-marginal: estudio piloto. *Anales de Salud Mental, 2*(1/2), 109-21.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*, 391-410.

Mathews, J. (1986). The professions of speech-language pathology and audiology. In G. Shames, & E. Wiig (Eds.), *Human communication disorders* (2nd ed., pp. 3-26). Columbus, OH: Charles E. Merrill Publishing.

May, R. B., Masson, M. J., & Hunter, M. A. (1990). *Application of statistics in behavioral research*. New York: Harper & Row.

Mazer, B. L., Korner-Bitensky, N., & Sofer, S. (1998). Predicting ability to drive after stroke. *Archive of Physical and Medicine Rehabilitation, 79*, 743-750.

McDermott, P. A. (1999). National scales of differential learning behaviors among American children and adolescents. *School Psychology Review, 28,* 280-291.

McDermott, P. A., Green, L. F., Francis, J. M., & Stott, D. H. (1999). *Learning Behaviors Scale.* Philadelphia: Edumetric and Clinical Science.

Meisels, S. J., Marsden, D. B., Wiske, M. S., & Henderson, L. W. (1997). *Early Screening Inventory - Revised: Examiner's manual*. Ann Arbor, Michigan: Rebus, Inc.

Meléndez, C. & Morocho, G. (2007). Aplicación de la Prueba de Predicción Lectora (PPL): Aspectos teóricos y elaboración de un baremo hecho en el Perú. *Investigación Educativa, 11*(19), 79-88.

Mercer, C.D. (1991). *Dificultades de aprendizaje* (Vol. I). Barcelona. CEAC.

Merino, C. & Angulo (2008). Test de Percepción Visual No Motor. *Avances en Medición, 6*(1), 163-168.

Merino, C. & Benites, L. (2011). Evaluación de la confiabilidad del Sistema Cualitativo de Calificación para la versión modificada del Test Gestáltico de Bender. *Universitas Psychologica, 10*(1), 237-249.

Merino, C. & Lautenschlager, G. (2003). Comparación estadística de la confiabilidad alfa de Cronbach: Aplicaciones en la medición educacional. *Revista de Psicología – Universidad de Chile, 12*(2), 129 – 139.

Merino, C. & Muñoz, P. (2007). Impacto socioeconómico sobre los puntajes de una batería multidimensional de aptitudes en niños preescolares. *Interdisciplinaria, 24*(2), 161-184.

Merino, C. & Sotelo, L. (2007). *Propiedades psicométricas del Test de Percepción Visual No Motor*. Documento no publicado.

Merino, C. (2008a). Un análisis de la varianza común, específica y de error en una prueba de aptitud cognitiva en preescolares: Efectos de la comprensión verbal. Datos no publicados.

Merino, C. (2008b). Una escala estandarizada de detección del bajo rendimiento escolar en niños de primer y segundo grado. Datos no publicados.

Merino, C. (2008c). Una exploración psicométrica del TONI-2 en niños escolares. Datos no publicados.

Merino, C. (2009d). Un análisis no paramétrico de ítems de la Prueba Gestáltica del Bender Modificada para estudiantes de primaria. *Liberabit, 15*(2), 83-94.

Merino, C. (En prensa).Validez incremental del Test Gestáltico de Bender Modificado, en niños de primer grado. *Avances en Psicología Latinoamericana*.

Merino, C., Angulo, M. & DeRoma, V. (2008). El Dibujo de la Figura Humana para la Estimación Intelectual (DAP:IQ): Estudio psicométrico preliminar en niños preescolares y escolares. Documento no publicado.

Meza, A. (1988). Psicología educacional en el Perú. *Psicología y Sociedad, 1*, 17-94.

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*, 156-166.

Milicic, N. & Schmidt, S. (1980). *Prueba de Precálculo*. Santiago: Galdoc.

Ministerio de Salud (2005). *Manual de Crecimiento y Desarrollo*. Lima: Autor.

Mora, J. A. (1999). *Batería Evaluadora de las Habilidades Necesarias para el Aprendizaje de la Lectura y Escritura (BEHNALE)*. Madrid: TEA.

Morote, W. C. & Robles, P. (1978). *La influencia del desarrollo socio-emocional en la madurez para el aprendizaje de la lectura-escritura en niños de 4 y 5 años de bajo status socio-económico*. Tesis para optar el Título en Psicología, Facultad de Psicología. Lima: Universidad Nacional Mayor de San Marcos.

Morse, D. T. (1999). MINSIZE2: A computer program for determining effect size and minimum sample size for statistical significance for univariate, multivariate, and nonparametric test. *Educational and Psychological Measurement, 59*(3), 518-531.

Morse, D. T.. & Merino, C. (2010). Mínimo tamaño muestral y magnitud del efecto para pruebas estadísticas inferenciales: Un programa para tu estimación. En preparación.

Muñiz, J. (1998). *Teoría Clásica de los Test (5ta ed.)*. Madrid. Pirámide

Muter, V. (2000). Screening for early reading failure. In N. Badian (Ed.), *Prediction and prevention of reading* failure (pp. 1-30). Parkton, MD: York Press.

Nieto, M. (2007). *Prueba Pedagógica – Primer Grado 2008*. Documento no publicado. I. E. María Inmaculada. Chorrillos, Lima: Autor.

Nunnally, J. C., & Bernstein, I. J. (1995). *Teoría psicométrica*. México, D. F.: McGraw-Hill.

Olivera, C., & Bernardo de la Cruz, V. (1987) *Instrumento de evaluación de las funciones intelectuales básicas del Aprestamiento*. Lima: INIDE.

Olivera, C., & Porras, M. (1987). *Manual para la aplicación del instrumento de evaluación de las funciones intelectuales básicas del Aprestamiento*. Lima: INIDE.

Ortiz, M., Becerra, J., Vega, K., Sierra, P. & Cassiani, Y. (2010). Madurez para la lectoescritura en niños/as de instituciones con diferentes estratos socioeconómicos. *Psicogente, 13*(23), 107-130.

Palomino, L. & Reyes, C. (1976). *La evaluación de las operaciones intelectuales en educandos de 5 a 8 años en Lima Metropolitana: Manual de Pruebas*. Lima: INIDE

Pandolfi, A., Mathiesen, M., & Oliva, M. (1997). Estilo lingüístico en centros educativos para preescolares pobres. *Estudios Filológicos, 32*, 37-55.

Parsons, L., & Weinberg, S. L. (1993). The Sugar Scoring System for the Bender-Gestalt. *Perceptual and Motor Skills, 77,* 883-893.

Pett, M. A. (1997). *Nonparametric statistic for health care research: Statistics for small samples and unusual distributions*. Thousand Oaks: Sage.

Peugh, J. L. (2009). A practical guide to multilevel modeling. Journal of School Psychology, 48, 85-112.

Piedmont, R. L., & Hyland, M. E. (1993). Inter-item correlation frequency distribution analysis: A method for evaluating scale dimensionality. *Educational and Psychological Measurement, 53,* 369-378.

Piñasca, F. (1992). *Relación entre maduración visomotora y el aprendizaje de la lecto-escritura en un grupo de niños del primer grado de primaria*. Tesis para optar el Título en Psicología, Facultad de Psicología. Lima: Universidad de San Martín de Porres.

Piñasca, F. (1992). *Relación entre maduración visomotora y el aprendizaje de la lecto-escritura en un grupo de niños del primer grado de primaria*. Tesis para optar el Título en Psicología, Facultad de Psicología. Lima: Universidad de San Martín de Porres.

Polit, D. F., & Hungler, R. P. (1997). In*vestigación científica en ciencias de la salud*. México: McGraw Hill Interamericana.

Preacher, K. J., Rucker, D. D., McCallum, R. C., & Nicewander, W. A. (2005). Use of extreme group approach: A critical reexamination and new recommendations. *Psychological Methods, 10*(2), 178-192.

Prieto, G. & Muñiz, J. (2000). Un modelo para evaluar la calidad de los tests utilizados en España. *Papeles del Psicólogo, 77,* 65-71.

Quiroz, L. (1976). *Valor pronóstico de los Test ABC y su correlación con el aprendizaje de la lecto-escritura*. Tesis de Bachiller en Psicología, Facultad de Psicología, Universidad Inca Garcilaso de la Vega. Lima: UPIGV.

Rathvon, N. (2004). *Early Reading Assessment: A Practitioner's Handbook*. New York: Guilford.

Rencher, A. C. (2002). *Methods of Multivariate Analysis* (2nd. Ed.). New York, NY: Wiley.

Reynolds, C. R. (2000). Methods for detecting and evaluating cultural bias in neuropsychological tests. In L. S. Fletcher-Janzen, L. S. Strickland & C. R. Reynolds (Eds.), *Handbook of Cross-Cultural Neuropsychology* (pp. 249-285). New York: Kluwer Academiz / Plenum Publishers.

Reynolds, C. R., & Hickman, J. A. (2004). *Draw-A-Person Intellectual Ability Test for Children, Adolescents, and Adults (DAP: IQ)*. Austin, TX: Pro-Ed.

Rosnow, R. L., & Rosenthal, R. (2008). *Begining behavioral research* (6th ed.). Upper Saddle River, NJ: Pearson/Prentice-Hall.

Rubio, J. (1992). *Estandarización de la Prueba de Funciones Básicas*. Instituto de Investigaciones Psicológicas, Facultad de Psicología: Universidad Nacional Mayor de San Marcos.

Salvessen. K. A., & Undheim, J. O. (1994). Screening for learning disabilities. *Journal of Learning Disabilities*, *27*(1), 60-66.

Sánchez, E. (1995). *Influencia de las nociones básicas de aprestamiento en niños con alto y bajo rendimiento matemático de primer y segundo grado de primaria pertenecientes a colegios estatales*. Tesis para optar el Título en Psicología, Facultad de Psicología. Lima: Universidad Ricardo Palma.

Sánchez, H., & Salazar, V. (1985). *Instrumento de evaluación de las nociones básicas del Aprestamiento*. Lima: INIDE.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.

Schapira, I., Aspres, N., Benitez, A., & Galindo, A. (1998, septiembre). *Hallazgos en los dibujos de prematuros de 3 a 5 años de vida*. Ponencia presentada en el VI Congreso Argentino de Perinatologia, Sociedad Argentina de Pediatría y Sociedad de Obstetricia y Ginecología de Buenos Aires, 10 al 12 de septiembre de 1998, Buenos Aires.

Schonhaut, L. Maggiolo, M, Barbieri, Z., Rojas, P., & Salgado, A. (2007). Dificultades de lenguaje en preescolares: Concordancia entre el test TEPSI y la evaluación fonoaudiológica. *Revista Chilena de Pediatría, 78*(4), 369-375.

Shepard, L. A. (1997). Children not ready to learn? The invalidity of school readiness testing. *Psychology in the Schools, 34,* 85–97.

Simmos, J. O. (1988). Fluharty Preschool Speech and Language Test: Analysis of construct validity. *Journal of Speech and Hearing Disorders, 53,* 168-174.

Sink, C.A., & Stroh, H. R (*2006*). *Practical significance*: *Use* of effect sizes in school counseling research. *Professional School Counseling, 9*, 401-411

Solan, H. A., Mozlin, R., & Rumpf, D. A. (1985). Selected perceptual norms and their relationship to reading in kindergarten and primary grades. *Journal of the American Optometric Association, 56*(6), 458-466.

Stanley, J. C. (1971). Reliability. In R. L. Thorndike (Eds), *Educational measurement (2nd ed.*, pp. 356-442*)*, Washington, D. C.: American Council On Education.

Su, C-Y., Charm, J-J., Chen, H-M., Su, C-J., Chien, T-H., Huang, M-H. (2000). Perceptual differences between stroke patients with cerebral infarction and intracerebral hemorrhage. *Archive of Physical and Medicine Rehabilitation, 81*, 706-714.

Sugar, F. R. (1995). *Sugar Scoring System for the Bender-Gestalt Test*. MA: Educator Publishing Service.

Terrones, A., Solís, M., Canudas, R., & Díaz, D. (1994). Vocabulario y nivel de comprensión pre-escolar y sin dicha experiencia educativa. *Revista Latinoamericana de Psicología, 26*(1), 83-95.

Thomas, A. (2001) School readinnes. En W. Edward Craighead & Charles B. Nemeroff (Eds.), *The Corsini Encyclopedia of Psychology and Behavioral Science* (3rd ed., Vol. 4, pp. 1455-1456). New York. Wiley & Sons.

Thorndike, R. L. (1989). *Psicometría aplicada*. México, D.F: Limusa.

Toledo, A. (1996) *Validez predictiva de los test del Dibujo de la Figura Humana de Goodenough-Harris, el Test Libre de Cultura de Cattell (Escala 1) y el Test de Vocabulario de Figuras de Peabody, con respecto al rendimiento en los cursos de Lenguaje y Matemáticas en alumnos de Primaria*. Tesis para optar la Licenciatura en Psicología. Lima: Universidad Inca Garcilaso de la Vega.

VanDerHeyden, A. M., Witt, J., Naquien, G., & Noell, G. (2001). The reliability and validity of Curriculum-based Measurement readiness probes for kindergarten students. *School Psychology Review, 30*(3), 363-382.

Velarde, E. (2004). La conciencia fonológica como zona de desarrollo próximo: Tesis revolucionaria de Luis Bravo Valdivieso. *Educación, 1*(2), 83-94.

Vera, E. (1988). *Estudio sobre la madurez para el aprendizaje de la lectura y escritura, a través de la Prueba 5 – 6, Forma B en niños de Centros de Educación Inicial*

particulares y estatales. Tesis para optar el Título en Psicología, Facultad de Psicología. Lima: Universidad de San Martín de Porres.

Visser, P. S., Krosnick, J. A., & Lavrakas, P. J. (2000). Survey research. In C. M. Judd & H. Reis (Eds.), *Research methods in social psychology* (pp. 223-252). NY: Cambridge University Press.

Wasserman, J. D., & Bracken, B. A. (2003). Psychometric considerations of assessment procedures. In J. Graham and J. Naglieri (Eds*). Handbook of assessment psychology (pp. 43 – 66).* New York: Wiley.

Werts, C. E., Rock, D. A., Linn, R. L., & Joreskog, K. G. (1978). A general method for estimating the reliability of a composite. *Educational and Psychological Measurement, 38*, 933-938.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: problems and remedies. In Hoyle, R. H. (Ed.), *Structural equation modeling: concepts, issues and applications* (pp. 56–75). Newbury Park, CA: Sage.

Whiteman T. A. (1987). *The PASS first grade screening test: Statistical analysis and predictive validity*. Unpublished manuscript, Department of Human Development, Bryn Mawr Collegue.

Wiener, I. B. (2003). The assessment process. In J. Graham & J. Naglieri (Eds.), *Handbook of assessment psychology* (Vol. 10, pp. 3-26). New York: Wiley.

Woodburn S. S. & Boschini C. (1993). Predictive validity of a spanish-language adapted version of the Anton Brenner Developmental Gestalt Test of School Readiness. *Perceptual and Motor skills, 76*(1), 315-318.

Woodburn, S. S. & Boschini, C. (1995). *Los problemas de aprendizaje en niños*. Heredia, San José: EU-EUNA.

Woodburn, S. S. & Boschini, C. (1997). *Los problemas de aprendizaje en niños*. Heredia, San José: EU-EUNA.

Woodburn, S. S., Boschini, C., Zeledón, M. & Fenández, H. (2004) *El desarrollo gestalt en los niños: Prueba Preescolar de Configuraciones*. Heredia, Costa Rica: UNA.

Worrell, F. C., Vandiver, B. J., & Watkins, M. W. (2001). Construct validity of the Learning Behavior Scale with an independent sample of students. *Psychology in the Schools, 38*(3), 207-215.

Yang, H. (2007). Normal curve. En N. J. Salkind (Eds.), *Encyclopedia of Measurement and Statistics* (Vol. 2, pp. 690-695). Thousand Oaks, CA: Sage.

York, C. D., & Cermak, S. A. (1995). Visual perception and praxis in adults after stroke. *American Journal of Occupational Therapy, 49*(6), 543-550.

Zarzosa, M. (2003). *El Programa de Lectura de Nivel 1 sobre la comprensión de lectura en niños en niños que cursan el 3er grado de primaria de nivel socioeconómico medio y bajo*. Tesis para optar el Título de Psicólogo, Facultad de Psicología. Lima: Universidad Nacional Mayor de San Marcos.

Zeledón, M. (1991). *Confiabilidad y objetividad de la prueda Gestalt de Anton Brenner, adaptada, con niños costarricenses entre 5 1. 2 y 7 años de edad*. Tesis de Licenciatura, Escuela de Ciencias del Deporte, Heredia, Costa Rica.

Zensano, W. (1995). *Relación del nivel intelectual con la madurez del aprendizaje para la lectura en un grupo de niños del primer grado del C.E. Nro. 2063 del distrito del Rimac*. Tesis para optar el Título en Psicología, Facultad de Psicología. Lima: Universidad Nacional Federico Villareal.

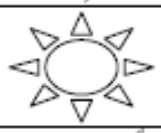
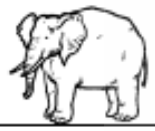
Zimmerman, D. W., Zumbo, B. D., & Williams, R. H. (2003). Bias in estimation and hypothesis testing of correlation. *Psicológica, 24*, 133-158.

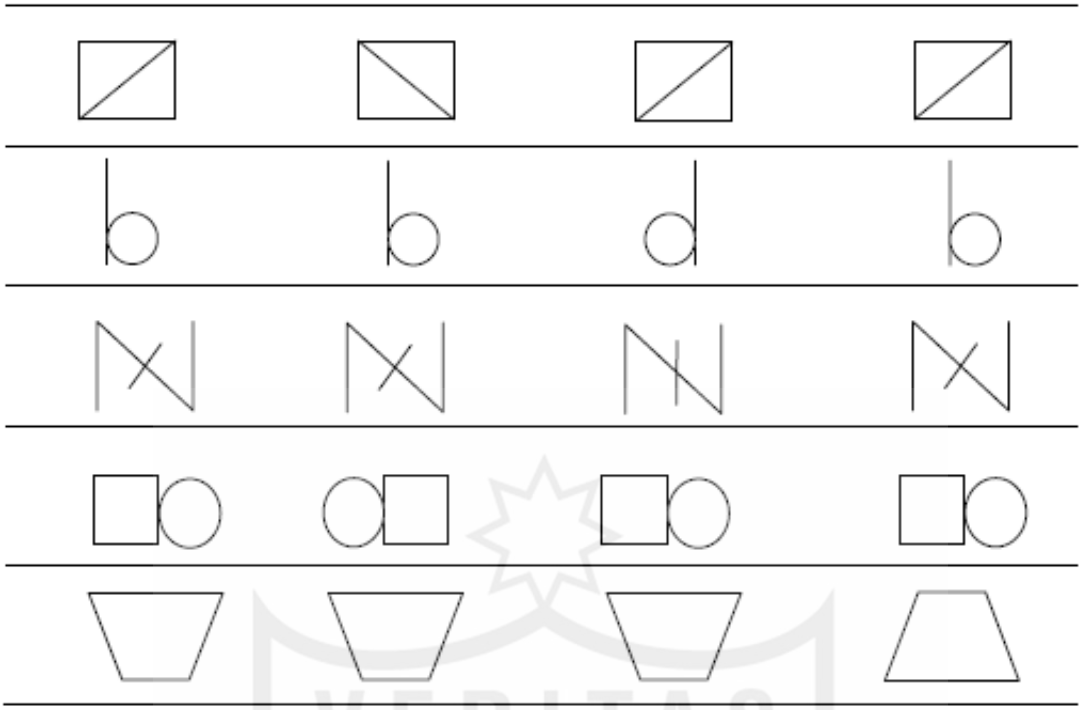
ANEXO

ANEXO A

Batería de Despistaje para Primer Grado (BDPG)

o	p	g	b	j	c	q
i	c	u	m	n	b	d
L	A	D	S	B	C	P
I	U	D	N	V	M	S
po	pu	pi	pa	pe		
mi	me	mo	pa	mu		
ae	eo	io	eu	ea		
ua	ui	iu	ou	ia		

	m k c l
	a c s d
	c e o a
   	
   	
   	
	a j o b
	e a o i
	d i m a
	mamá mam____
	____rbol
	____jo
	aut____
	lun____



mono	mona	mopo	mano	mono
sopa	sopo	sopa	soqa	sapa
bota	bola	boca	bota	dota
dado	dada	dobo	doda	dado
dumbo	dumdo	dumbo	bumdo	bumbo
lemna	lemna	lomna	lenma	lemno

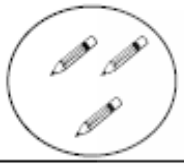
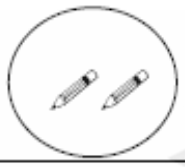
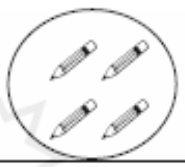
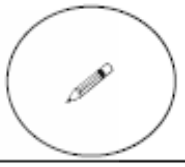
nesa	meca	mase	mesa
mape	mopi	napi	maqi
ana	ama	asa	ata

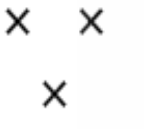
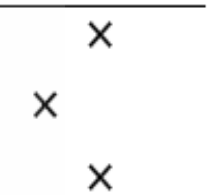
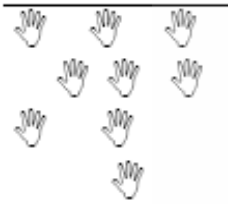
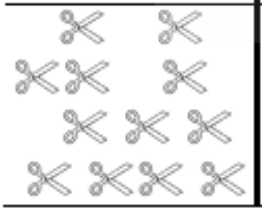
mercado - mercado	careta - caseta
marido - manido	zapalo - zapalo
mama - mana	terrijo - terijo
miboro - miboro	peqito - pepito
calzapo - calzapo	amigo - amigo

L	B	3	7	T	6	g
C	2	Z	=	G	4	h

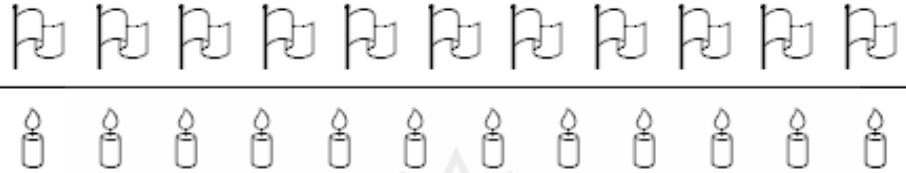




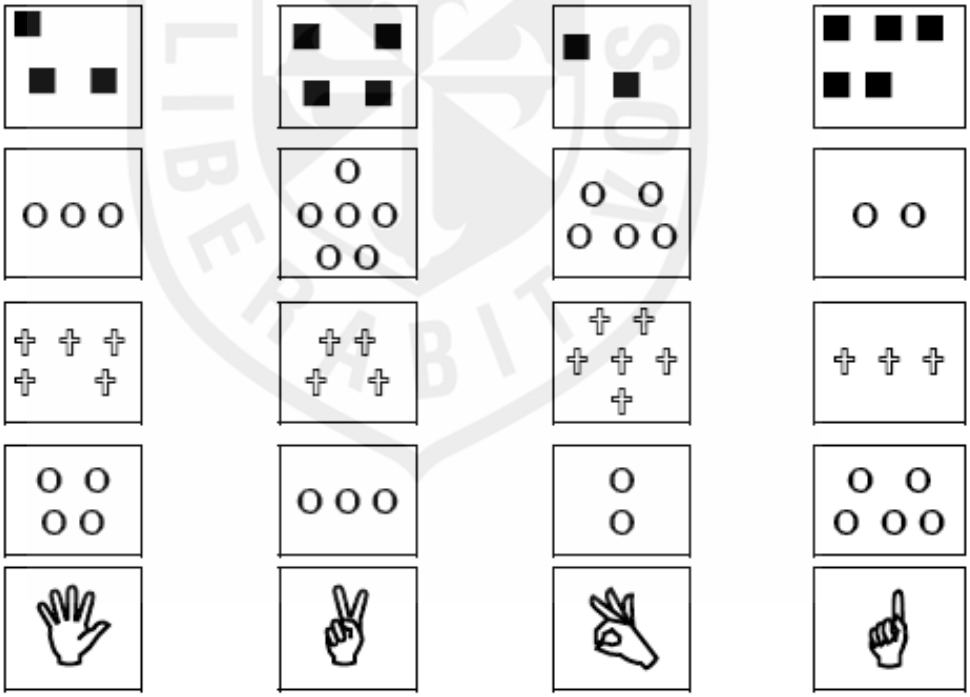
				
				
	9	5	8	4
	10	13	12	8

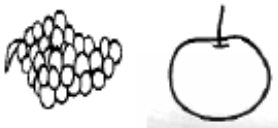
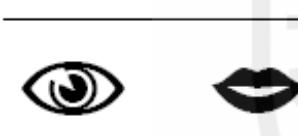
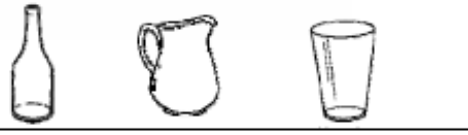
7	
5	

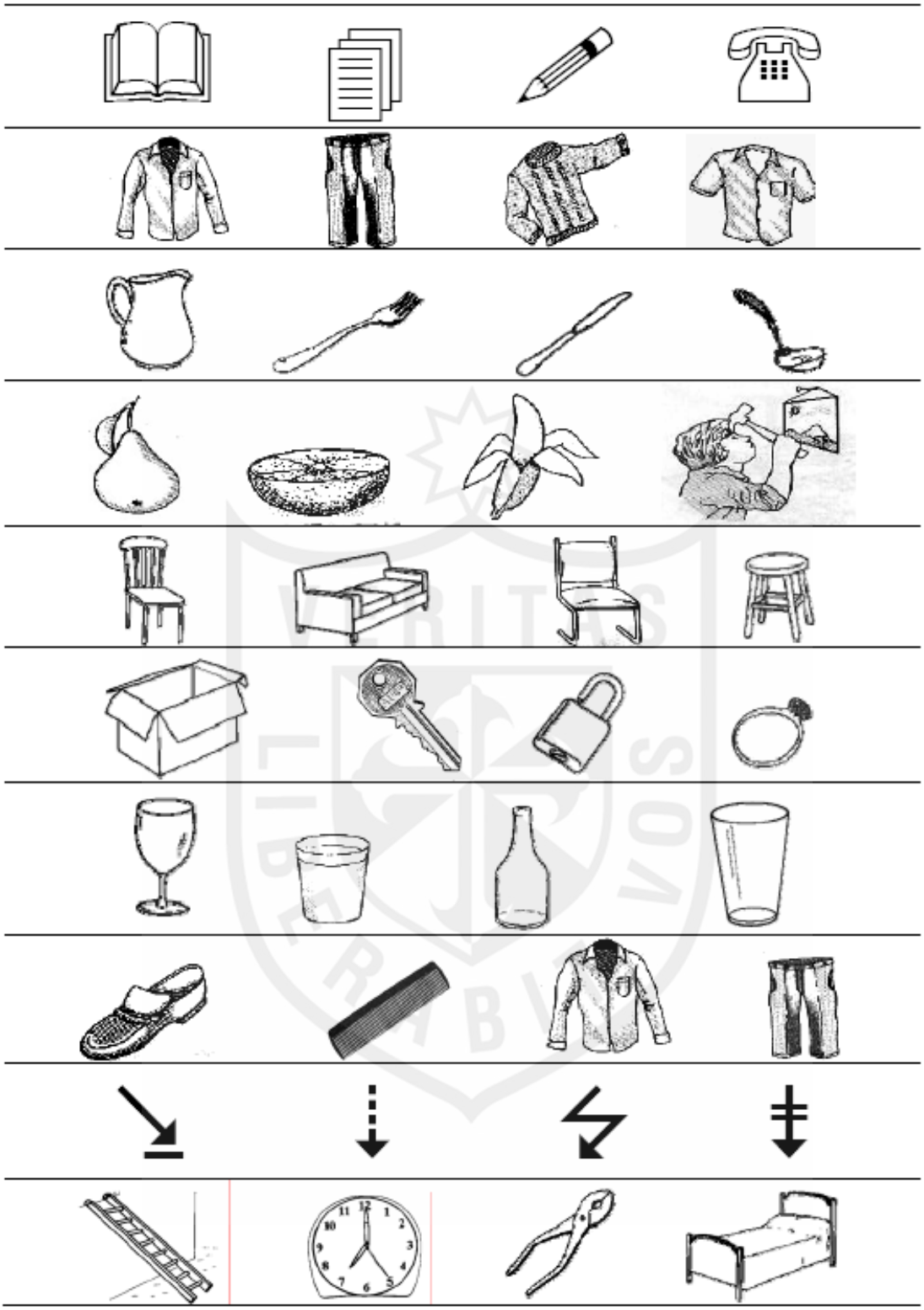


11 ___ 13 ___ ___ 16

5 ___ 7 ___ 9 ___



 _____	
	
 _____	
 _____	
	
 _____	
 _____	
3 5 _____	m 2 p
 _____	
 _____	



Manual de aplicación

Indicaciones generales

- Se puede repetir hasta tres veces la instrucción de cada ítem.
- Individualmente, se puede repetir las indicaciones si parece no haber entendido. En el apoyo individual, no dar ayuda para hallar la respuesta correcta y se debe repetir las indicaciones hasta dos veces más.
- Sin un niño señala su respuesta y la enseña al examinador, pero no la marca, insistir que debe marcar su respuesta:

  Marca el que [tarea]….

- Si el niño continua sin marcar, esperando una respuesta aprobatoria del examinador, decir:

  ¿Cuál es? [esperar su respuesta]..... *márcalo.*

- El examinador debe mostrarse neutral ante niño, respecto a lo correcto o incorrecto de la respuesta. Si el niño(a) insiste en una respuesta errónea, dejar que lo marque. Es más probable que el niño no sepa la respuesta y no que no haya entendido la instrucción
- Luego, para las siguientes ocasiones, indicar:

  No me digas cuál es,... sólo marca tu respuesta...

- Se da esta indicación sea si el niño señalado una opción correcta o incorrecta. Algunas veces, el niño insiste en esperar una respuesta del examinador; el examinador, entonces, debe abreviar la frase anterior con una indicación manual señalando que debe hacerlo solo.
- Después que los niños voltean la hoja para las siguientes tareas, supervisar que todos estén la hoja de estímulos apropiada.

- Los subtests se aplican en el orden en que aparecen impresas.

Instrucciones específicas

"Vamos a hacer un juego con las letras."

"Escuchen bien lo que haremos."

"Pongan su dedo en esta parte [Mostar y verificar que todos estén apuntando]

PAGINA 1

8. Aquí escribe tu nombre	
9. Aquí en este espacio, escriban las vocales	*Señalar el espacio*
10. Entre estas letras, busquen y marquen la letra **p** ... Bajen su dedito... 11. Entre estas letras, busquen y marquen la letra **m** 12. Entre estas letras, busquen y marquen la letra **c** 13. Entre estas letras, busquen y marquen la letra **s**	*Señalar la línea de letras* *Señalar la línea de letras* *Repetir instrucciones para las letras C y S*
14. Ahora, en este espacio, marquen lo que les voy a decir: **PA**	*Señalar el espacio* *Repetir el estímulo hasta dos veces para todos*
15. Ahora, debajo marquen **MI**	*Señalar el espacio* *Repetir el estímulo hasta dos veces para todos*
16. Ahora marquen: **EA**	*Señalar el espacio* *Repetir el estímulo hasta dos veces para todos*
17. Ahora marquen **IU**	*Señalar el espacio* *Repetir el estímulo hasta dos veces para todos*
18. Ahora escriban lo que les voy a decir: **PI**	*Señalar el espacio* *Repetir el estímulo hasta dos veces para todos*
19. Ahora escriban lo que les voy a decir: **MO**	*Señalar el espacio* *Repetir el estímulo hasta dos veces para todos*

Ahora vamos hacer un juego diferente. Me van a decir cuántos sonidos hay en esta palabra: ZAPATO. Escuchen de nuevo. ZA *[Pausa]* PA *[Pausa]* TO. Son tres sonidos y debo dibujar tres aspas. Ahora, díganme cuántos sonidos hay en esta palabra: DEDO ....Muy bien..... Hay dos sonidos y entonces debo dibujar 2 aspas. 20. Ahora, en este espacio van a dibujar con aspas la cantidad de sonidos que tiene la palabra que les voy a decir: Atención **CASITA**	*Se pronuncia las sílabas una por segundo. Dibujar en la pizarra un aspa apareándola temporalmente con cada sonido pronunciado.* *Para el segundo ejemplo, pronunciar normalmente y escribir en la pizarra las aspas.*
21. Ahora escuchen **PALO**	
22. Ahora escuchen **CAMINITO**	

PAGINA 2

Item 23

¿Ven la figura de la mano? ..... pongan su dedito sobre la figura de la **MANO**.... al lado en esta parte hay varias letras. Marquen la letra con la que comienza la palabra MMMANO. Vean como lo digo... MMMANO.... Marquen.	*Señalar la línea de letras* *Asegurar que los niños vean al examinador pronunciar la palabra*

Item 24

Ahora ..... pongan su dedito sobre la figura del **SOL**. Al lado en esta parte hay varias letras. Marquen la letra con la que comienza la palabra SOL. Vean como lo digo... SSSOL.... Marquen.	*Señalar la línea de letras* *Asegurar que los niños vean al examinador pronunciar la palabra*

Item 25

Ahora ..... pongan su dedito sobre la figura del **AVION**. Al lado en esta parte hay varias letras. Marquen la letra con la que comienza la palabra AVION. Vean como lo digo... AAAVION.... Marquen.	*Señalar la línea de letras* *Asegurar que los niños vean al examinador pronunciar la palabra*

Ahora, haremos otro juego. Vamos a escuchar como terminan las palabras.

Item 26

Van a marcar la palabra que es mas larga, la que me demoro más en decir. Por ejemplo..... CAMINO PAN..... ¿cuál es la más larga?.... muy bien es "camino" Aquí, tenemos **CABALLO**, **BEBE**, **GATO**, **OLLA** Marquen la palabra mas larga	*Desarrollar un ejemplo con una palabra de pronunciación larga y otra de pronunciación corta. Hacer gestos de la mano para indicar lo largo o corto de las palabras de ejemplo*

Item 27

Ahora, marquen la palabra más corta. Aquí tenemos: **ELEFANTE**, **OSO**, **TENEDOR**, **DINOSAURIO**,	*Hacer gesto de "corto"*

Item 28

Ahora, marquen la palabra más larga. Aquí tenemos: **PERRO**, **LIBRO**, **SILLA**, **MARIPOSA**	*No hacer ningún gesto*

Item 29

..... pongan su dedito sobre la figura del **SAPO**.... al costado hay varias letras Marquen la letra con la que termina la palabra SAPO.. Vean como lo digo... SAPOO.... Marquen.	*Señalar la línea de letras* *Asegurar que los niños vean al examinador pronunciar la palabra*

Item 30

..... pongan su dedito sobre la figura del **PIE**.... al costado hay varias letras. Marquen la letra con la que termina la palabra PIE. Vean como lo digo... PIEEE.... Marquen.	*Señalar la línea de letras* *Asegurar que los niños vean al examinador pronunciar la palabra*

Item 31

..... pongan su dedito sobre la figura de la **PELOTA**.... al costado en parte hay varias letras. Marquen la letra con la que termina la palabra PELOTA. Vean como lo digo... PELOTAA... Marquen.	*Señalar la línea de letras* *Asegurar que los niños vean al examinador pronunciar la palabra*

Item 32

Aquí hay una palabra,.... dice **MAMÁ**. Aquí al lado debe decir MAMÁ pero le falta la <u>última letra</u> aquí en esta rayita. Escriban la letra que falta para que podamos leer MAMA	*Copiar el ejemplo en la pizarra*

Item 33

Aquí vemos un **ARBOL**. Al lado debe decir ARBOL pero falta la primera letra de la palabra ARBOL, aquí en la rayita. Escriban la letra sobre la rayita para que la palabra diga AARBOL	*No decir qué letra es la faltante.*

Item 34

Aquí vemos un **OJO**. Al lado debe decir OJO pero falta la primera letra de la palabra OJO, aquí en la rayita. Escriban la letra sobre la rayita para que la palabra diga OOJO	*No decir qué letra es la faltante.*

Item 35

Aquí vemos un **AUTO**. Al lado debe decir AUTO pero falta la última letra aquí de la palabra AUTO en la rayita. Escriban la letra sobre la rayita para que la palabra diga AAUTO	*No decir qué letra es la faltante.*

Item 36

Aquí vemos un **LUNA**. Al lado debe decir LUNA pero falta la última letra de la palabra LUNA aquí en la rayita. Escriban la letra sobre la rayita para que la palabra diga LUNA	*No decir qué letra es la faltante.*

PAGINA 3

Items del 37 al 40

Aquí hay unas figuras. Son iguales pero una no es igual a las demás.... ¿Cuál será? Busquen y marquen la que es diferente, la que no es igual a las demás. Cuando terminen, hagan lo mismo con esta otra	*Señalar la línea de figuras* *Enfatizar que deben de marcar la que es diferente* *Señalar enfatizando que continúen con las demás luego de la primera.*

Item del 41 al 47

Aquí hay otro juego que vamos a hacer..... ¿Ven esta palabra?.... En esta parte busquen y marquen la palabra que es igual a esta. Cuando terminen, hagan lo mismo con esta otra, y sigan hasta terminar	*Señalar palabra MONO* *Señalar las palabras del panel derecho* *Señalar enfatizando que continúen con las demás luego de la primera.*

PAGINA 4

Item 48

Ahora, escuchen con atención. En esta parte, van a buscar y marcar la palabra que les voy a decir: **MESA**	*Señalar* *Pronunciar claramente y en fluidez normal. Se puede repetir una vez más.*

Item 49

Bajen su dedo. Ahora marquen la palabra: **MAPE**	*Señalar* *Pronunciar claramente y en fluidez normal. Se puede repetir una vez más.*

Item 50

Ahora, escuchan con atención. En esta parte, van a buscar y marcar la palabra que les voy a decir: **AMA** ... marquen.	*Señalar* *Pronunciar claramente y en fluidez normal. Se puede repetir una vez más.*

Item 51, 52, 53

51. Ahora, aquí haremos algo otra cosa: escribirás lo que les voy a decir. **PAPA**	*Señalar el espacio* *Se puede repetir una vez más al grupo o individualmente.*
52. Ahora, aquí escribe: **MAMÁ**	
53. Debajo escribe **MEMA**	

Item 54 al 63

Ahora, otro jueguito. Aquí hay algunas palabras que son iguales y otras que son diferentes. Miren aquí cómo lo hago. ¿Estas palabras son iguales? Si son iguales, entonces las marcan con un aspa. ¿Estas otras palabras son iguales? Muy bien, son diferentes,... entonces hago un círculo. Marquen con un aspa grande, así como lo hago yo, SOLO en las palabras que son iguales. Los que diferentes, encierro con un círculo Trabajen solos. Si son iguales hago un aspa; si son diferentes, hago un círculo	*Escribir en la pizarra dos palabras iguales y otras dos diferentes.* *Marcar ambas con un aspa grande cubriendo la extensión de las palabras.* *Enfatizar que solo deben de marcar rayitas que separan las palabras iguales; las que son diferentes debe encerrar con un círculo tal rayita.* *Supervisar luego de dar las instrucciones.* *Repetir varias veces y pausadamente durante el desarrollo de la prueba*

PAGINA 5

Item 64

Aquí marca todos los **números** que veas	*Señalar la línea correspondiente*

Item 65

Aquí también; marca todos los **números** que veas	*Señalar la línea correspondiente*

Item 66

¿Ven estos puntitos? Busquen y marquen en donde hay **MENOS** puntos	*Señalar la línea correspondiente*

Item 67

Aquí marquen donde hay **MAS** lápices	*Señalar la línea correspondiente*

Item 68

¿Ven estos relojes? Marquen el reloj que está en el **CENTRO**	*Señalar la línea correspondiente*

Item 69

Debajo marquen el **ULTIMO** anteojo	*Señalar la línea correspondiente*

Item 70

Este juego es diferente,... atención. Cuenten cuántas aspas hay aquí y busquen en este lado donde hay la **MISMA** cantidad de aspas,.... y lo marcan	*Señalar* *Enfatizar que deben marcar el que tiene la misma cantidad*

Item 71

Ahora aquí, donde están las caritas felices Cuenten cuántas caritas hay aquí y busquen en este lado el que tiene la **MISMA** cantidad de caritas,.... y lo marcan	*Señalar* *Enfatizar que deben marcar el que tiene la misma cantidad*

Item 72

Ahora aquí, donde están las manos. Cuenten cuántas manos hay, y aquí busquen y marquen el número de manos que han contado. Busca el número que has contado.	*Señalar* *Si es necesario, repetir:*

Item 73

Hagan lo mismo aquí con las tijeras. Cuenten cuántas tijeras hay, y aquí busquen y marquen el número de manos que han contado	*Señalar* *Si es necesario, repetir*

PAGINA 6

Item del 74 al 75

74. En este espacio, dibujen esta cantidad de bolitas	*Señalar el número 7 al referirse a la cantidad. No identificar el número al niño(a)*
75. Aquí también hagan lo mismo	*Señalar el número 5 al referirse a la cantidad. No identificar el número al niño(a)*

Ítems del 76 al 79

76. Escribe los números del 1 al 10	*Señalar el espacio indicado* *Si es necesario, repetir*
77. Aquí marca 7 banderitas	*Señalar el espacio indicado* *Si es necesario, repetir*
78. Debajo marca 5 velas	*Señalar el espacio indicado* *Si es necesario, repetir*
79. Dibuja 13 bolitas	*Señalar el espacio indicado* *Si es necesario, repetir*

Ítems del 80 al 81

Ahora miren aquí. Completen los números que faltan en las rayitas *Si parecen no comprender, indicar:* ¿Luego del 11 que sigue? … 12. Continúen.	*Señalar* *Para el ítem 81, usualmente los niños ya han entendido la tarea.* *No dar más ayuda*

Ítems del 82 al 86

82. Ahora, coloquen su dedo aquí. Escuchen. Tengo 5 cuadraditos y se me perdió uno. ¿Cuántos tengo? Marquen la cantidad que tengo.	*Señalar el lugar de colocación del dedo, al lado de la primera caja.* *Indicar que no digan la respuesta, sólo que la marquen.*
83. Ven estas bolitas. Marquen en que tiene menos de tres bolitas	*Se puede repetir la instrucción hasta dos veces*
85. Aquí marquen el que tiene más de cinco cruces	*Se puede repetir la instrucción hasta dos veces*
86. Aquí escuchen: Tengo 2 bolitas y me encuentro 1 más. ¿Cuántos tengo? Marquen la cantidad que tengo	*Se puede repetir la instrucción hasta dos veces.* *Indicar que no digan la respuesta, sólo que la marquen.*
86. Aquí marquen la mano que está enseñando más de 4 dedos	*Se puede repetir la instrucción hasta dos veces*

PAGINA 7

Ítems del 87 al 96

87. Aquí tenemos un grupo de figuritas,.. ¿Ven? Aquí hay dos figuras pero falta una en esta raya. Busquen aquí la figura que se parece a estas de aquí. *Si se observa que no comprenden, indica*r: Deben marcar la figura que se <u>parece</u> a los barquitos de aquí (señalar)	*Se repite hasta dos veces* *Señalar las opciones haciendo un escaneo en ellas.* *Esperar la respuesta y proseguir con la siguiente instrucción*
Sigan con los demás. Busquen y marquen la figura que va con este grupo.	*Se repite la instrucción cuando el niño tiene duda. No sugerir la respuesta correcta.*

PAGINA 8

97. Aquí marca lo que <u>no</u> se usa cuando escribimos	*Señalar el espacio indicado* *Si es necesario, repetir*
98. Bajen su dedo. Aquí marca lo que nos ponemos para abrigarnos cuando tenemos frío	*Señalar el espacio indicado* *Si es necesario, repetir*
99. Bajen su dedo. Aquí marca lo que usamos para llevar el agua	*Señalar el espacio indicado* *Si es necesario, repetir*
100. Bajen su dedo. Aquí marca lo que esta pelado	*Señalar el espacio indicado* *Si es necesario, repetir*
101. Bajen su dedo. Aquí marca el banco	*Señalar el espacio indicado* *Si es necesario, repetir*
102. Bajen su dedo. Aquí marca el anillo	*Señalar el espacio indicado* *Si es necesario, repetir*
103. Bajen su dedo. Aquí marca la copa	*Señalar el espacio indicado* *Si es necesario, repetir*
104. Bajen su dedo. Aquí marca lo que usamos para caminar	*Señalar el espacio indicado* *Si es necesario, repetir*
105. Bajen su dedo. Aquí marca lo que está doblado	*Señalar el espacio indicado* *Si es necesario, repetir*
106. Bajen su dedo. Aquí marca lo que tiene escalones	*Señalar el espacio indicado* *Si es necesario, repetir*

ANEXO B

Tabla A

Investigaciones publicadas y no publicadas sobre madurez escolar y variables relacionadas

Autor	Evidencias obtenidas	Variable	
		Explicativa	Respuesta
U. Inca Garcilaso de la Vega			
Albiragorta (1983)	Validez de constructo	Ajuste emocional	Madurez para la lecto-escritura
Toledo (1996)	Validez predictiva	Pruebas de habilidades	Rendimiento escolar en lenguaje y matemáticas
Gutierrez (1988)	Validez de constructo	Variables demográficas	Madurez para el aprendizaje
Quiroz (1976)	Validez predictiva	Test A.B.C	Aprendizaje de lecto-escritura
U. Nacional Federico Villareal			
Zensano (1995)	Validez de constructo	Nivel intelectual	Aprendizaje de la lectura (ABC)
Llactahuacchaca (1994)	Validez de constructo	Zona rural vs. Urbana	Percepción visual (Frosting)
Chumbipuma (1996)	Validez de constructo	Madurez social	Madurez para el aprendizaje de la lecto-escritura (5-6 B)
Correa (1977)	Validez de constructo	Variables demográficas	Prueba Jordan - Massey
U. Femenina del Sagrado Corazón			
Calderón (1991)	Validez predictiva	Examen de ingreso ad hoc	Rendimiento en 1er g.
Canales (1990)	Validez de constructo	Variables demográficas	Prueba de Pre-cálculo
Espinoza. & Matos (1991)	Validez de constructo	Grado escolar, edad	Percepción visual

Autor	Evidencias obtenidas	Variable	
		Explicativa	Respuesta
Espinoza, Piedra & Sotomarino (1995)	Estandarización	Prueba de Funciones Básicas	
U. Nacional Mayor de San Marcos			
Rubio (1992)	Estandarización	Prueba de Funciones Básicas	
Morote & Robles (1978)	Validez de constructo	Desarrollo socio-emocional	Madurez para la lectura
U. Ricardo Palma			
Sánchez (1995)	Validez predictiva	Aprestamiento	Rendimiento e 1er y 2do grado
U. de San Martín de Porres			
Alzadora (1981)	Validez de constructo	NSE	Desarrollo psicomotríz
Arraya (1992)	Validez predictiva	Coord. Visomotora	Problemas de aprendizaje
Piñasca (1992)	Validez predictiva	Maduración visomotora	Aprendizaje de lectura y escritura
Vera (1988)	Validez de constructo	Prueba 5-6	Madurez en centros estatales vs. privados
Investigaciones publicadas			
Majluf et al. (1971)	Validez de constructo	Variables demográficas	ABC Jordan y Massey

ANEXO C

Diferencia de grupos

Método Peters (1941) para la diferencia de grupos extremos.

Se usa la siguiente ecuación para obtener la correlación entre la variable que separa a los grupos (X) y el criterio (Y):

$$r = \frac{\left(\overline{Y_S} - \overline{Y_I}\right) P_S P_I}{DE_Y \left(P_U Z_I + P_I Z_U\right)}$$

$\overline{Y}_S$ = Puntaje promedio del grupo superior
$\overline{Y}_I$ = Puntaje promedio del grupo inferior
P_S = Proporción de sujetos en el grupo superior
P_I = Proporción de sujetos en el grupo inferior
Z_S = Ordenada de la curva normal correspondiente a PS
Z_I = Ordenada de la curva normal correspondiente a PI
DEY = Desviación estándar del criterio

Como se requiere la desviación estándar de distribución en la muestra del criterio Y (DE_Y), y solo se conservan los puntajes de Y en los grupos extremos, se aplicará una estimación de la desviación estándar total combinada asumiendo distribución normal del criterio (Garret, 1971):

$$DE_{comb} = \sqrt{\frac{N_S \left(DE_S^2 + d_S^2\right) + N_I \left(DE_I^2 + d_I^2\right)}{N_{comb}}}$$

N_s = Número de sujetos en el grupo superior
N_i = número de sujetos en el grupo inferior
DE_s^2 = varianza del grupo superior en Y (criterio)
DE_i^2 = varianza del grupo inferior en Y
d_s^2 = distancia cuadrada de la media del grupo superior a la media ponderada en Y
d_i^2 = distancia cuadrada de la media del grupo inferior a la media ponderada en Y
N_{comp} = número de sujetos total

La *r* calculada por el método Peters (1941) requiere ser ajustada por el efecto de la segmentación de la variable o restricción del rango (Preacher, 2007):

$$r_{xyT} = \sqrt{1 - \left(\frac{DE_c^2}{DE_T^2} \right) \left(1 - r_{xyC}^2 \right)}$$

r_{xyT} = Correlacion estimada en el rango completo

r_{xyC} = correlación desde la muestra segmentada

DE_T^2 = varianza en el rango completo de la variable Y

DE_c^2 = varianza de la variable Y segmentada