Banco Interamericano de Desarrollo

DOCUMENTO DE TRABAJO DEL BID Nº IDB-WP-723

Validez concurrente y viabilidad de pruebas cortas comúnmente usadas para medir el desarrollo infantil temprano en estudios a gran escala: Metodología y resultados

Marta Rubio-Codina
María Caridad Araujo
Orazio Attanasio
Sally Grantham-McGregor

Banco Interamericano de Desarrollo
División de Protección Social y Salud

Agosto 2016

BID
Banco Interamericano de Desarrollo

Validez concurrente y viabilidad de pruebas cortas comúnmente usadas para medir el desarrollo infantil temprano en estudios a gran escala: Metodología y resultados

Marta Rubio-Codina
María Caridad Araujo
Orazio Attanasio
Sally Grantham-McGregor

Agosto 2016

Catalogación en la fuente proporcionada por la
Biblioteca Felipe Herrera del
Banco Interamericano de Desarrollo
Validez concurrente y viabilidad de pruebas cortas comúnmente usadas para medir el desarrollo infantil temprano en estudios a gran escala: metodología y resultados / Marta Rubio-Codina, María Caridad Araujo, Orazio, Attanasio, Sally Grantham-McGregor.
p. cm. — (Documento de trabajo del BID ; 723)
Incluye referencias bibliográficas.
1. Child development-Colombia-Evaluation. 2. Early childhood education-Colombia-Evaluation. 3. Educational tests and measurements-Colombia. I. Rubio-Codina, Marta. II. Araujo, María Caridad. III. Attanasio, Orazio P. IV. Grantham-McGregor, Sally M. V. Banco Interamericano de Desarrollo. División de Protección Social y Salud. VI. Serie.
IDB-WP-723

http://www.iadb.org

scl-sph@iadb.org

www.iadb.org/SocialProtection

Validez concurrente y viabilidad de pruebas cortas comúnmente usadas para medir el desarrollo infantil temprano en estudios a gran escala: Metodología y resultados

Marta Rubio-Codina[1,2], María Caridad Araujo[1], Orazio Attanasio[2,3], Sally Grantham-McGregor[4]

Resumen[†]

Medir el desarrollo infantil temprano (DIT) mediante pruebas estándar de diagnóstico en estudios a gran escala resulta difícil y costoso. Por esto, con frecuencia se utilizan pruebas de tamizaje multidimensionales y pruebas que miden un solo dominio del desarrollo como alternativas (nos referimos a ellas como 'pruebas cortas'). La validez de las pruebas cortas en estos contextos es desconocida. En este estudio, analizamos la viabilidad, confiabilidad y validez concurrente de algunas de estas pruebas cortas de uso habitual, incluyendo tres pruebas de tamizaje multidimensionales—los Cuestionarios de Edades y Etapas, la Prueba de Tamizaje del Desarrollo de Denver, el Inventario del Desarrollo de Battelle—y dos pruebas que miden un solo dominio del desarrollo—la versión abreviada de MacArthur-Bates y los Hitos del Desarrollo Motor—en 1.311 niños y niñas de 6 a 42 meses en Bogotá, Colombia. Comparamos las puntuaciones obtenidas en estas pruebas cortas con las de las Escalas Bayley de Desarrollo Infantil, prueba que consideramos nuestro patrón de referencia. El Bayley se implementó en un centro por psicólogos, mientras que las pruebas cortas se realizaron en los hogares por encuestadores, tal y como se administrarían en una encuesta de hogar. La validez concurrente de las escalas cognitivas, de lenguaje y de motricidad fina de las pruebas multidimensionales con la escala correspondiente en el Bayley es baja en niños menores de 19 meses, si bien aumenta con la edad convirtiéndose en moderada a alta a partir de los 30 meses. En cambio, la concurrencia de las escalas de motricidad gruesa es alta en menores de 19 meses y disminuye a partir de esta edad. Con respecto a las pruebas que miden un solo dominio del desarrollo, los Hitos del Desarrollo Motor presentan un alto grado de validez con la motricidad gruesa en niños menores de 16 meses, y las escalas de lenguaje expresivo de la versión abreviada de MacArthur-Bates muestran una correlación moderada con el lenguaje en menores de 30 meses. Por dominio del desarrollo, la motricidad gruesa presenta el grado de validez más alto en menores de 19 meses, mientras que en niños de más de 19 meses, el desarrollo del lenguaje es el dominio que obtiene concurrencias mayores.

Palabras clave: evaluación del desarrollo, pruebas de diagnóstico, pruebas de tamizaje, validez concurrente, cognición, lenguaje, desarrollo motor, primera infancia, estudios a gran escala, países de ingresos bajos y medios.

Códigos JEL: J1, I1, I2, I3

[1] División de Protección Social y Salud, Banco Interamericano de Desarrollo, Washington DC, EE.UU.
[2] Centro para la Evaluación de Políticas de Desarrollo, Instituto de Estudios Fiscales, Londres, RU.
[3] Departamento de Economía, University College London, Londres, RU.
[4] Instituto de Salud Infantil, University College London, Londres, RU.

[†] Este documento presenta en mayor detalle el diseño y la metodología de trabajo del estudio: Rubio-Codina M, Araujo MC, Attanasio O, Muñoz P, Grantham-McGregor S (2016) Concurrent Validity and Feasibility of Short Tests Currently Used to Measure Early Childhood Development in Large Scale Studies. PLoS ONE 11(8): e0160962. Las opiniones, hallazgos y recomendaciones presentes en este documento reflejan las opiniones de los autores y no necesariamente las del BID, sus Directores Ejecutivos, ni las de los países que ellos representan. La recolección de datos fue financiada por el Fondo RG-T1907 del Banco Interamericano de Desarrollo (BID). Rubio-Codina agradece el financiamiento parcial del programa Early Career Fellowship ECF/2008/0170 de Leverhulme Trust. El trabajo de investigación de Attanasio fue financiado parcialmente por el Advanced Grants 249612 del Consejo Europeo de Investigación (ERC) y por el Professorial Fellowship ES/K010700/1 del Consejo Económico y Social de Investigación (ESRC). Agradecemos a todas las familias que participaron en el estudio y a BibloRed, Jardines Sociales del Distrito de Bogotá y aeioTU por prestarnos sus instalaciones para realizar las pruebas. Extendemos nuestro agradecimiento a todos los encuestadores y evaluadores, capacitadores (Pablo Muñoz, Mara Minski y Natalia Varela) y coordinadores de campo (Belén Gómez, Juan Fernando Trujillo y Hanner Sánchez). Agradecemos también enormemente a Stefano Banfi, Ludvig Sinander y Camila Soares por su contribución como asistentes de investigación. Todos los errores deben atribuirse solo a los autores.

1. Introducción

Estudios recientes han demostrado la importancia de los primeros años de vida para el desarrollo del cerebro, el desarrollo cognitivo, del lenguaje y socioemocional y, en términos más generales, para la formación del capital humano (Luby 2015; Heckman 2007). Estudios longitudinales muestran que la adversidad en la primera infancia tiene efectos de largo plazo en el desarrollo del niño (Walker *et al.* 2011) y se estima que alrededor de 200 millones de niños y niñas menores de 5 años en países de ingresos bajos y medios no logran alcanzar su potencial de desarrollo (Grantham-McGregor *et al.* 2007). Las intervenciones en la primera infancia pueden generar impactos en varios indicadores en la edad adulta (Walker *et al.* 2011; Gertler *et al.* 2014; Campbell *et al.* 2014), y existe un compromiso a nivel mundial cada vez mayor para implementar este tipo de intervenciones a escala en países de ingresos bajos y medios a fin de promover el desarrollo de niñas y niños en condiciones de vulnerabilidad. Los Objetivos del Desarrollo Sostenible (ODS), por ejemplo, tienen como una de sus metas que para el 2030 "todas los niños y niñas tengan acceso a un desarrollo infantil temprano de calidad, servicios de cuidado y educación preescolar a fin de que estén preparados para la enseñanza primaria" (ODS 4.2) (UN General Assembly 2015).

No obstante, la implementación de intervenciones para el desarrollo infantil temprano (DIT) se ve limitada por la falta de instrumentos de medición del desarrollo infantil confiables y válidos que permitan recolectar datos de manera costo-efectiva en muestras de gran tamaño (Engle *et al*. 2007; Frongillo *et al.* 2014). Estos instrumentos son imprescindibles tanto para medir niveles de desarrollo a nivel poblacional como también para monitorear y evaluar la eficacia de intervenciones, lo que nos puede brindar información acerca de cómo mejorarlas. Además, estos instrumentos son fundamentales en la estimación de modelos de acumulación de capital humano, los cuales contribuyen a un mejor entendimiento acerca del proceso de desarrollo de habilidades a lo largo del ciclo de la vida, incluyendo el rol de los padres y cuánto invierten en sus hijos durante los primeros años (Heckman 2007; Attanasio 2015). La necesidad de contar con instrumentos que midan los resultados del DIT es particularmente apremiante en niños menores de 3 años. Por esto mismo, es imperativo identificar, de entre las ya disponibles, herramientas de medición válidas y confiables para la evaluación del desarrollo infantil en muestras de gran tamaño que puedan ser recogidas en el contexto de una encuesta de hogar (es decir, 'a escala').

Las pruebas de diagnóstico multidimensionales, como por ejemplo las Escalas Bayley de Desarrollo Infantil (Bayley 1969; Bayley 2006), son consideradas el patrón de referencia para medir los resultados del DIT en menores de 3 años y medio (Frongillo *et al.* 2014; Fernald *et al.* 2009; Fernandes *et al.* 2014). Además, esta prueba ha demostrado ser sensible a diferencias en indicadores de desarrollo como resultado de intervenciones de DIT en diversos contextos (Hamadani *et al.* 2006; Nahar *et al.* 2009; Attanasio *et al.* 2014). Sin embargo, la administración de esta prueba es muy larga y requiere profesionales capacitados que trabajen en entornos controlados; los materiales y otros costos de administración (pago por niño, por ejemplo) son muy altos, y encontrar profesionales que puedan administrarla en el idioma local es todo un desafío. Por otro lado, para adaptar estas pruebas a otros idiomas y contextos culturales se necesitan profesionales capacitados, tiempo y recursos financieros. Por estos motivos, el Bayley y otras pruebas de diagnóstico similares a menudo resultan inviables para aplicarse a escala.

Como alternativa, en estudios a gran escala y evaluaciones de impacto es cada vez más frecuente el uso de pruebas de tamizaje, diseñadas específicamente para detectar niños en riesgo de desarrollo, y pruebas que miden un dominio específico del desarrollo (p. ej. dominio del lenguaje) (Fernald *et al.* 2012; Macours, Schady y Vakis 2012; Fernald e Hidrobo 2011). A pesar de que estas pruebas no fueron diseñadas para este propósito y en muchas ocasiones no han sido validadas, ni estandarizadas localmente, están ganando popularidad. Esto se debe a que son pruebas más cortas, más económicas y más fáciles de administrar, en parte porque con relativa frecuencia se basan en varios ítems obtenidos por reporte materno y porque en muchas ocasiones son administradas por encuestadores en el hogar.

No obstante, poco se conoce acerca de su validez cuando se administran a escala, no para tamizaje sino para medir niveles de desarrollo del niño a lo largo de toda la distribución del desarrollo, tanto para fines de investigación como para obtener indicadores de desarrollo a nivel poblacional. Dos excepciones recientes son los estudios de Hamadani y colegas en zonas rurales de Bangladesh (Hamadani *et al.* 2013; Hamadani *et al.* 2010). Los autores encontraron correlaciones moderadas entre los reportes maternos acerca de la edad en que el niño logra ciertos hitos en motricidad gruesa—principalmente caminar y pararse solo—y el Índice de Desarrollo Psicomotor del Bayley-II (PDI, por sus siglas en inglés). También hallaron asociaciones—bajas pero significativas—con el Índice de Desarrollo Mental (MDI, en inglés) a los 18 meses y con el coeficiente intelectual a los 5 años (Hamadani *et al.* 2013). De modo similar, una prueba de lenguaje en niños de entre 12 y 18 meses, desarrollada localmente a partir de los Inventarios de Desarrollo de Habilidades Comunicativas MacArthur-Bates (Fenson *et al.* 2002) y administrada mediante reportes maternos, presentó una validez concurrente moderada con el MDI del Bayley-II y una validez predictiva aceptable con el coeficiente intelectual a los 5 años (Hamadani *et al.* 2010). Es interesante observar que los reportes maternos acerca de la edad en la que el niño comienza a caminar solo y de los niveles de vocabulario que presenta, resultaron ser tan predictivos del desarrollo motor o del coeficiente intelectual a los 64 meses como el PDI y MDI del Bayley-II, respectivamente.

Más recientemente, se han desarrollado nuevas pruebas de diagnóstico multidimensionales para utilizar en niños de 24 meses o más en países de ingresos bajos y medios. Algunos ejemplos son el proyecto INTERGROWTH-21st (Fernandes *et al.* 2014) para la medición del desarrollo infantil a los 24 meses o la Escala Engle, desarrollada por el Banco Interamericano de Desarrollo en el marco del Proyecto Regional de Indicadores de Desarrollo Infantil (PRIDI) (Verdisco *et al.* 2009) y destinada a niños y niñas de entre 24 y 59 meses. La ONG Save the Children desarrolló la herramienta Evaluación Internacional del Desarrollo Infantil y el Aprendizaje Temprano (*International Development and Early Learning Assessment,* IDELA) para medir el desarrollo y aprendizaje en niños de entre 3,5 y 6 años de edad, a través de indicadores de prelectura y prematemática, entre otros (Wolf *et al.* 2015). Asimismo, la Institución Brookings, bajo la Comisión Especial sobre Métricas del Aprendizaje, ha dirigido a un grupo de trabajo en el desarrollo de un instrumento que mida la calidad de los entornos de aprendizaje y las habilidades preacadémicas y socioemocionales en niños de 3 a 5 años. Esta iniciativa se conoce como el Proyecto de Medición de la Calidad y los Resultados del Aprendizaje Temprano (*Measuring Early Learning Quality and Outcomes,* MELQO).[1] No obstante, es importante destacar que estas iniciativas no incluyen a niños menores de 2 años y que muchas de las nuevas pruebas diseñadas siguen siendo demasiado largas como para ser utilizadas a escala.

[1] http://www.brookings.edu/about/centers/universal-education/learning-metrics-task-force-2/melqo

Este estudio tiene como objeto contribuir al debate actual sobre la medición de los resultados del DIT, que rápidamente está atrayendo la atención de investigadores y profesionales en instituciones de diversa índole. La investigación está diseñada para establecer en qué medida las pruebas de tamizaje multidimensionales y las pruebas que miden un solo dominio del desarrollo que se incluyeron en el estudio (de aquí en adelante 'pruebas cortas') son alternativas válidas y viables para evaluar el desarrollo de niños pequeños a escala. Más concretamente, nuestro objetivo es determinar la confiabilidad del test-retest, la consistencia interna y la validez concurrente de cinco pruebas cortas, administradas en condiciones similares a las de una encuesta (por encuestadores en el hogar del niño), para medir los niveles de desarrollo en una muestra de niños de entre 6 y 42 meses en Bogotá, Colombia. También analizamos los tiempos y costos de administración relativos de cada una de estas pruebas.

Las pruebas cortas que se seleccionaron han sido habitualmente en estudios a escala. Se limitó la cantidad de pruebas para evitar que el niño se canse demasiado o se afecte su bienestar. En particular, el estudio incluye las siguientes pruebas cortas: tres pruebas de tamizaje multidimensionales—los Cuestionarios de Edades y Etapas (*Ages and Stages Questionnaires*, tercera edición, ASQ-3) (Squires *et al.* 2009), la Prueba de Tamizaje del Desarrollo de Denver (*Denver Developmental Screening Test*, segunda edición, Denver-II) (Frankenburg *et al.* 1990; Frankenburg *et al.* 1992) y el Inventario del Desarrollo de Battelle (*Battelle Developmental Inventory screener*, segunda edición, BDI-2) (Newborg 2005); y dos pruebas que miden un solo dominio del desarrollo—la lista de vocabulario de la versión abreviada de los Inventarios I y II del Desarrollo de Habilidades Comunicativas MacArthur-Bates (*MacArthur-Bates Communicative Development Inventories I and II Short Forms*, SFI y SFII) (Jackson-Maldonado *et al.* 2003; Jackson-Maldonado, Marchman y Fernald 2012) y los Hitos del Desarrollo Motor Grueso de la Organización Mundial de la Salud (*WHO Motor Milestones*, WHO-Motor) (WHO Multicentre Growth Reference Study Group 2006; Wijnhoven et al. 2004). Las últimas dos pruebas comparten numerosas similitudes con las pruebas utilizadas en los estudios bangladesís mencionados anteriormente (Hamadani *et al.* 2010; Hamadani *et al.* 2013) y se incluyeron debido a que su capacitación y administración requieren mucho menos tiempo y son menos costosas.

Para calcular la validez concurrente se compararon las puntuaciones de desarrollo que los niños obtuvieron en estas pruebas cortas con las puntuaciones obtenidas en las Escalas Bayley de Desarrollo Infantil (*Bayley Scales of Infant and Toddler Development*,tercera edición, Bayley-III) (Bayley 2006). Por tratarse de nuestro patrón de referencia, el Bayley-III se administró en condiciones óptimas, es decir, en centros específicamente habilitados para la administración de la prueba y por psicólogos capacitados. Por otro lado, todas las pruebas cortas se administraron tal y como habitualmente se haría en el contexto de una encuesta: en el hogar del niño y por encuestadores no especializados en el área del desarrollo infantil y sin experiencia previa en la administración de pruebas, pero que sí recibieron una capacitación rigurosa. Esto es importante ya que permite abordar las preguntas clave de esta investigación.

Analizamos la validez concurrente de las pruebas cortas a través del Bayley-III por edad del niño y por dominio del desarrollo, centrándonos en las áreas cognitiva, el lenguaje receptivo y expresivo y la motricidad fina y gruesa. A pesar de que consideramos que el desarrollo socioemocional es un dominio del desarrollo importante y recogimos los datos de la escala, no lo hemos incluido en este análisis. Esto se debe en parte a que las escalas de conducta adaptativa y de desarrollo personal-social de las pruebas cortas miden constructos algo distintos a los de la escala socioemocional del Bayley-III, lo que limita la capacidad de

comparación. Por otro lado, la escala socioemocional se recoge a través de los reportes del cuidador principal y en consecuencia es un procedimiento ya de por sí relativamente rápido y fácil. Volveremos a este punto en la próxima sección. También analizamos la validez concurrente según el estatus socioeconómico del hogar para investigar qué pruebas cortas son las más adecuadas para administrarse en las familias más desfavorecidas, que por lo general reciben una menor educación y a quienes es más probable que se dirijan los programas sociales del gobierno.

Es importante tener en cuenta que este estudio no se diseñó con el fin de determinar la sensibilidad o especificidad que tienen estas pruebas de tamizaje para identificar a niños con alto riesgo de rezago o retraso en el desarrollo. De hecho, el número de niños en riesgo de padecer estos problemas en el desarrollo en la muestra es muy reducido como para que se pueda llevar a cabo tal análisis. Nuestro interés es más bien analizar la capacidad de las pruebas cortas para medir el nivel de desarrollo de los niños en la población estudiada, una población que representa a los grupos sociales de clase baja y media-baja de una ciudad grande, típica de un país de ingresos bajos y medios en América Latina. El objetivo es identificar, de entre los ya disponibles, aquellos instrumentos confiables y fáciles de administrar que sirvan para evaluar a niños pequeños en estudios a gran escala y en entornos distintos de los contextos para los que fueron creados, guiando de esta forma la decisión respecto de qué instrumento utilizar en investigaciones (en la evaluación de intervenciones, por ejemplo) o en la medición del desarrollo a nivel poblacional.

Este artículo está organizado de la siguiente manera. La Sección 2 describe el diseño del estudio y las estrategias de recolección de datos. Además, incluye una descripción sobre las pruebas que se administraron y la muestra final de análisis. La Sección 3 presenta la estrategia empleada para el análisis empírico. La Sección 4 expone los resultados y por último, sobre la base del análisis previo, la Sección 5 profundiza en el debate acerca de qué pruebas conviene utilizar a escala y concluye.

1. Diseño del estudio y recolección de datos

1.1. Participantes y estrategia de recolección de datos

Bogotá está dividida en seis estratos socioeconómicos (sectores) según su ubicación y la calidad de las viviendas y de la infraestructura. Este estudio incluyó una muestra representativa de niños y niñas de entre 6 y 42 meses seleccionados de forma aleatoria de entre los tres estratos más pobres de la ciudad, estratificando por edad y sector. Estos tres estratos conforman el 85% de la población de Bogotá y comprenden hogares de ingresos bajos y medios.[2] En un principio habíamos incluido al estrato 4 (de ingresos medios) en el diseño del estudio, sin embargo se acabó excluyéndolo debido a que fue muy difícil contactar a las familias de este estrato y obtener su consentimiento de participación. Esto se atribuye a que muchas de estas familias viven en apartamentos y recintos confinados de acceso restringido. En efecto, la falta de confianza fue uno de los principales motivos por los que un gran número de personas en el Estrato 4 no quisieron participar. En cuanto a las edades de los niños que se incluyeron en el estudio, establecimos los seis meses como edad mínima, puesto que las mediciones en niños más pequeños presentan una capacidad predictiva baja respecto del desarrollo futuro y también debido a restricciones de

[2] Los barrios de los estratos 1 y 2 se consideran por lo general pobres, mientras que aquellos del estrato 3 se consideran de clase media-baja. Sin embargo, es necesario aclarar que hay bastante heterogeneidad en las características socioeconómicas de los hogares dentro de cada estrato, en especial en aquellos barrios que se han creado recientemente (véase Rubio-Codina *et al*. 2015).

presupuesto. Determinamos la edad máxima según lo que establece el Bayley-III que está diseñado para evaluar a niños de hasta 42 meses de edad.[3]

Los datos se recogieron entre marzo y agosto de 2011. La muestra incluye niños muy pequeños y debía ser representativa por estrato y estar balanceada por grupo etario. Asimismo, era importante asegurarse de que los niños de todas las edades y estratos fueran evaluados en proporciones similares durante todo el período de campo con el fin de minimizar los efectos estacionales y los efectos relacionados con la curva de aprendizaje o la fatiga del evaluador. No obstante, al iniciar el estudio no teníamos acceso a registros administrativos que contaran tanto con las fechas de nacimiento de los niños como con sus domicilios, incluyendo el estrato, como para saber dónde encontrarlos. Así, el levantamiento de la muestra implicó un desafío logístico considerable que requirió seguir una estrategia rigurosamente definida de antemano que implementamos de manera estricta en tres etapas que se desarrollaron simultáneamente en cada barrio.

En primera instancia, los barrios (y las manzanas dentro de cada barrio) se seleccionaron aleatoriamente ponderando por el porcentaje de mujeres en edad fértil (diseño probabilístico). Una vez seleccionados los barrios, visitamos puerta a puerta todos los hogares a fin de identificar a las familias con niños de entre 6 y 42 meses. Esto se llevó a cabo por un equipo de encuestadores dedicados exclusivamente a identificar la muestra del estudio. Se excluyó a niños con dificultades en el aprendizaje (un niño) y mellizos (un par) por razones prácticas. Del mismo modo, en aquellos hogares en que hubiera más de un niño entre las edades comprendidas en el estudio (cuatro casos) se incluyó solo a uno de ellos aleatoriamente. El resto de los niños elegibles para participar en el estudio se estratificaron por edad, y entre ellos se seleccionó al 80% (por manzana y grupo etario) de manera aleatoria para que forme parte de la muestra.

El siguiente paso consistía en asignar todos los niños identificados e incluidos en el estudio en una misma manzana a una de las ocho encuestadoras—capacitadas pero no especializadas en el área del DIT—quienes visitaron cada hogar a fin de administrar las pruebas cortas y una encuesta. Esta encuesta incluía información básica acerca del nivel socioeconómico del hogar (composición demográfica, nivel educativo y situación laboral de cada miembro, características de la vivienda y los bienes y activos que poseen); la historia de salud del niño (peso al nacer, edad gestacional, entre otros); información acerca de la forma de cuidado formal (p. ej. en institución de cuidado) e informal (p. ej. con familiares) que ha recibido el niño, así como también la calidad del entorno familiar medida a través de los indicadores del cuidado familiar (*Family Care Indicators*, FCI) de UNICEF (Frongillo, Sywulka y Kariger 2003). En términos específicos, registramos por observación la cantidad de libros por adulto que había en el hogar, los diarios/revistas y los tipos de juguetes con los que los niños jugaban usualmente; y por reporte del cuidador principal, las actividades de juego que realizaron los niños con sus padres durante la semana previa a la encuesta.

Durante la etapa final, las psicólogas capacitadas (evaluadoras) administraron el Bayley-III en las bibliotecas públicas o en centros desarrollo infantil que se encontraran cerca del hogar del niño.[4] Esto garantizó que todas las evaluaciones del Bayley-III se llevaran a cabo en entornos similares, y se cumpliera así con los requisitos de administración de la prueba

[3] Es muy probable que si se incluía a niños más pequeños en el estudio, esto hubiese limitado la disponibilidad de participar de las familias o la posibilidad de hacerlo dado que el Bayley-III y las mediciones antropométricas se realizaron fuera de los hogares de los niños.

[4] Esto se llevó a cabo gracias a la colaboración de la red local de bibliotecas públicas BibloRed y los centros públicos de atención a la infancia Jardines Sociales. Por habernos prestado sus instalaciones, ofrecimos al personal y a los padres de los centros y bibliotecas talleres sobre prácticas de crianza y habilidades parentales.

(silencio, buena iluminación, espacio adecuado, ventilación), lo que permitió además que el niño se concentrara y se aprovechara mejor el tiempo de administración. En promedio, el Bayley-III se administró entre cinco y seis días después de que se realizaran las pruebas cortas (78% en el plazo de una semana y 94% en el plazo de dos semanas). Además, las evaluadoras desconocían los resultados que los niños habían obtenido en dichas pruebas. Luego de haber finalizado el Bayley-III, la evaluadora recogió los datos relativos a la talla y peso de la madre y del niño siguiendo las directrices de la OMS (WHO 1983). Como muestra de agradecimiento por haber participado en el estudio, se regaló a los niños que fueron evaluados un set de libros de ilustraciones y suplementos nutricionales (vitaminas y minerales) para consumo diario durante tres meses. Del mismo modo, se entregó a la madre información sobre el desempeño del niño en la prueba, un set de folletos para padres y COP 10.000 (alrededor de USD 5,6) para cubrir los gastos de transporte del hogar al centro correspondiente.

Con el fin de incluir un mayor número de pruebas en el estudio, sin que esto resultara un proceso agotador para los niños y sus familias, se asignó cada niño, de manera aleatoria, a una de las dos baterías de pruebas cortas que se crearon. La Batería A incluyó la prueba ASQ-3, el Denver-II y para niños de entre 8 y 30 meses el SFI o el SFII, dependiendo de la edad. Por otro lado, la Batería B comprendía el BDI-2 y para niños de 6 a 15 meses el WHO-Motor. Las pruebas cortas se administraron en el orden en que han sido mencionadas dentro de cada batería y luego de haberse completado la primera sección de la encuesta en el hogar, es decir, una vez entablada la relación con el cuidador principal. El tiempo de administración de ambas baterías fue muy similar y la duración total de la visita domiciliaria (encuesta en el hogar + pruebas cortas) no superó las dos horas o dos horas y media. Esto permitió que el encuestador completara entre dos o tres visitas domiciliarias por día (el promedio de visitas diarias aumentaba a medida que se avanzaba con la recolección de datos). Del mismo modo, cada evaluador administraba entre dos o tres Bayley-III por día. Entre un 2,5% y un 5% de las sesiones, ya sea en el hogar o en el centro, tuvieron que ser reprogramadas debido a que el niño se encontraba enfermo o estaba demasiado inquieto o quisquilloso como para que fuera posible realizarle las pruebas.

Todas las pruebas (pruebas cortas y Bayley-III) se llevaron a cabo en presencia del cuidador principal—la madre en un 85-89% de los casos y el padre un 3-5% de las veces. En los casos restantes, el cuidador principal del niño era en la mayoría de ocasiones otro pariente. El cuidador respondía los ítems de las pruebas cuando así se requería. Por este motivo y para asegurarnos de que el niño o niña estuviera acompañado de alguien familiar y que le transmitiera confianza y seguridad durante la administración de las pruebas, solicitamos que la persona que acompañara al niño fuera mayor de 15 años y que por lo general pasara al menos cinco horas por día cuidándolo, un mínimo de cinco días a la semana.

El Gráfico 1 presenta un resumen del diseño y de las etapas del estudio, en el que se enumeran todas las pruebas administradas por batería y se indica el número de participantes en cada etapa y prueba. Se monitorearon las edades y estratos de todos los niños en la muestra durante todo el proceso a fin de garantizar una muestra final balanceada. Además, la recolección de datos se organizó de tal forma que, durante los seis meses que duraron las actividades, todas las encuestadoras y evaluadoras evaluaran consistentemente a una cantidad similar de niños por estrato y grupo etario. Esto fue importante para disminuir posibles sesgos en la medición que pueden surgir por varios de los siguientes factores: (i) estatus socioeconómico del niño (p. ej. el evaluador otorga distintas puntuaciones según el entorno de donde provienen los niños para compensar las

situaciones de desventaja que percibe); (ii) edad del niño (p. ej. al evaluador le resulta más fácil evaluar a niños mayores); (iii) estacionalidad (p. ej. las mediciones son menos precisas cuando se realizan de forma apresurada cerca de períodos de vacaciones o fines de semana largos), y (iv) efectos relacionados con la fatiga o la curva de aprendizaje del evaluador (p. ej. las pruebas son más confiables cuando se administran durante el período intermedio de la recolección de datos, una vez que el evaluador ya ha practicado lo suficiente, pero cuando todavía no está tan cansado de haber realizado la misma prueba una y otra vez). En otras palabras, queríamos evitar que los patrones por edad o por sector socioeconómico que se observen en los datos estuviesen relacionados con alguna de las posibles fuentes de sesgo enumeradas.

El comité de ética del Instituto de Ortopedia Infantil Roosevelt en Bogotá revisó los protocolos de estudio y los consideró completamente acordes con las prácticas éticas requeridas. Los padres de los niños participantes firmaron el consentimiento informado en su nombre. Para mayor información acerca de los procedimientos de selección de la muestra y recolección de datos véase Rubio-Codina *et al.* (2015) y Rubio-Codina, Attanasio y Grantham-McGregor (2016).

1.2. Instrumentos de medición del DIT

Las dos primeras columnas en la Tabla 1 contienen las pruebas y las escalas (i.e. dominios del desarrollo) que se han administrado en este estudio. Es importante observar que tres de las pruebas cortas—concretamente, el ASQ-3, el Denver-II y el BDI-2—abarcan múltiples dimensiones, mientras que el WHO-Motor y el SFI y SFII son pruebas que miden un solo dominio del desarrollo: motricidad gruesa y lenguaje, respectivamente. Al lado de cada escala, se indica entre paréntesis el total de ítems de la prueba y entre corchetes el promedio de ítems evaluados por niño en el estudio. Para las pruebas con puntos de partida y techo, el número de ítems administrados por niño se establece en función de su edad y sus habilidades y, por consiguiente, estos dos valores no coinciden.

Las siguientes dos columnas representan el rango etario que abarcan las pruebas y las edades en que se administraron en el estudio. Nótese que no todas las pruebas cubren el rango etario del estudio en su totalidad. El resto de las columnas presentan otras características de las pruebas como el costo de los materiales (sin incluir los gastos de envío y aduana) y los costos de administración por niño; el tiempo de administración de acuerdo con lo informado por la editorial de las pruebas, y el tiempo de administración y capacitación promedio en el estudio.[5] Las últimas dos columnas contienen la valoración de los capacitadores sobre el grado de dificultad para capacitar y administrar cada prueba.

A pesar de que la mayoría de las pruebas estaban disponibles en español, algunas tuvieron que ser traducidas total o parcialmente. Asimismo, luego de pilotear las versiones (oficiales) en español o sus traducciones se consideró necesario realizar algunas modificaciones en la redacción y estilo para reflejar mejor el español de Colombia; así como contextualizar algunas imágenes. En el Apéndice I enumeramos las modificaciones que hemos realizado y los sitios web de las editoriales. A continuación ofrecemos una descripción detallada de cada prueba.

[5] El tiempo total de administración de las pruebas fue registrado por el capacitador durante las evaluaciones que fueron supervisadas (alrededor del 5% de la muestra).

1.2.1. Prueba de referencia: Escalas de Desarrollo Infantil de Bayley, tercera edición (Bayley-III)

El Bayley-III (Bayley 2006) es una prueba de diagnóstico que consiste en las siguientes escalas:

(i) Escala cognitiva. Se basa principalmente en respuestas no verbales del niño y mide los procesos de aprendizaje, la capacidad de resolver problemas, la atención, la habilidad para contar objetos y clasificarlos, y las habilidades para jugar, entre otros constructos.
(ii) Escala de lenguaje y comunicación. Dentro de este dominio se encuentran las subescalas de lenguaje receptivo y expresivo. La primera subescala mide la capacidad del niño de comprender los distintos estímulos, las palabras o las instrucciones en el entorno. La segunda evalúa el desarrollo del lenguaje a través de las vocalizaciones, el uso de palabras y la construcción de oraciones.
(iii) Escala motora. Incluye la subescala de motricidad fina que mide la coordinación manos-dedos y manos-ojos, y la subescala de motricidad gruesa que mide el control del niño sobre su cuerpo y las habilidades para mover torso y extremidades.
(iv) Escala socioemocional. Se mide a través de la Gráfica de Desarrollo Socioemocional de Greenspan (*Greenspan Social-Emotional Growth Chart*, Greenspan 2004) y evalúa los principales hitos del desarrollo socioemocional, como la autorregulación, la atención, la habilidad del niño de relacionarse e interactuar con familiares y desconocidos, entre otros aspectos temperamentales y sociales.
(v) Escala de conducta adaptativa. Se mide a través del Formulario del Padre/Cuidador Principal del Sistema de Evaluación de la Conducta Adaptativa (*Parent/Primary Caregiver Form of the Adaptive Behavior Assessment System*, segunda edición, ABAS-II) (Harrison y Oakland 2003) y consiste en diez subescalas que evalúan las habilidades funcionales diarias de niños de 0 a 5 años de edad.[6]

Las escalas se administran y se puntúan de forma independiente, lo que produce evaluaciones específicas para cada dominio. Las escalas cognitiva, de lenguaje y motora se evalúan a través de la observación directa de las habilidades del niño en varios ítems que están ordenados en un grado ascendente de dificultad. Criterios de inicio (base) y parada (techo) determinan los ítems de la prueba que realiza cada niño. Por cada ítem que el niño realiza correctamente recibe un puntaje de 1, si no logra ejecutarlo el puntaje es 0. El puntaje bruto es la suma de respuestas correctas, incluyendo los ítems anteriores al punto de inicio (base).

Como se mencionó anteriormente, el foco de este estudio es el desarrollo cognitivo, de lenguaje y motor. La escala socioemocional comprende 35 preguntas de cinco puntos cada una que debe responder el cuidador, por lo que su administración es de por sí bastante rápida y fácil. Sin embargo, al centrarse más en la medición de aspectos de autocuidado y autodirección, las escalas de desarrollo personal-social y de conducta adaptativa de las pruebas cortas no son muy comparables con la escala socioemocional del Bayley-III y por esto no la hemos incluido en el análisis. En cuanto a la escala de conducta adaptativa, solo se recogieron los datos de dos subescalas del ABAS-II en una submuestra de niños debido a restricciones de tiempo y a que ni las edades de los niños ni su contexto eran siempre los adecuados para evaluar muchos de los ítems en el resto de subescalas. Por esto, estas

[6] Las diez áreas que abarca son las siguientes: comunicación, utilización de recursos comunitarios, habilidades preacadémicas funcionales, vida en el hogar, salud y seguridad, ocio, autocuidado, autodirección, social y motora.

escalas, también administradas mediante reportes del cuidador, tampoco han sido incluidas en el análisis.

Para la aplicación del Bayley-III se necesitan profesionales en el área de desarrollo infantil, como psicólogos y educadores, que hayan recibido una capacitación rigurosa. El tiempo de administración es de entre 30 y 90 minutos, según la edad del niño. En nuestro caso, la evaluación de las escalas cognitiva, de lenguaje, motora y socioemocional tuvo una duración promedio de 83 minutos, con un rango de 40 a 150 minutos dependiendo de las características del niño (edad, interés, atención, etc.). De hecho, el tiempo de administración aumenta considerablemente con la edad en niños menores de 24 meses y luego se mantiene estable. Así, el tiempo promedio de aplicación fue de 77 minutos para los niños menores de 24 meses y de 93 minutos para los más grandes.

La prueba completa cuesta USD 1.050 e incluye un cuaderno de estímulos, un libro de imágenes, un set de objetos manipulativos (muñecos, pelotas, patos de goma, tablero con formas geométricas, rompecabezas, bloques, etc.), un manual técnico y un manual de administración, y 25 hojas (cuadernos) de respuesta para cada escala. Se pueden comprar hojas de respuesta adicionales para cada escala o para varias escalas en conjunto. Cada una de estas tiene un valor de USD 9,34 si incluye todas las escalas, o de USD 5,02 para las escalas cognitiva, de lenguaje y motora. El precio unitario de cada hoja de respuesta extra equivale al costo de administración por niño, ya que la editorial exige la compra de una hoja por cada niño evaluado. En un estudio a gran escala, con tamaños de muestra grandes, esto puede representar un costo de administración prohibitivo. Adicionalmente, se requieren otros materiales para realizar la prueba que no están incluidos en el paquete, como tijeras, cinta adhesiva, lápices, un cronómetro y un set de escalones de dimensiones específicas necesarios para evaluar la motricidad gruesa. Únicamente pueden comprar la prueba aquellos profesionales que trabajen en esta área y que posean títulos en educación superior (p. ej. doctorado en psicología, en educación o en disciplinas estrechamente relacionadas), o aquellas personas que posean acreditación o sean miembros de alguna organización profesional específica y estén capacitadas para administrar las pruebas e interpretar los datos recolectados.

El Bayley-III está disponible en español desde mediados de 2015. Por ello, para el presente estudio, hubo que traducir la versión en inglés de los manuales y de las hojas de respuesta al español de Colombia y luego traducirlos de nuevo al inglés (para controlar la confiabilidad de la traducción).

1.2.2. Pruebas cortas por validar en la Batería A

1.2.2.1. Cuestionarios de Edades y Etapas, tercera edición (ASQ-3)

El ASQ-3 (Squires *et al.* 2009) es un instrumento de tamizaje destinado a niños de 1 a 66 meses. Está compuesto por 21 cuestionarios edad-específicos que debe responder el cuidador principal. Cada cuestionario evalúa el desarrollo del niño en cinco dominios (escalas) —resolución de problemas (o escala cognitiva), comunicación, motricidad fina, motricidad gruesa y personal-social. Cada escala a su vez contiene seis ítems.

Como prueba de tamizaje, el ASQ-3 está diseñado para identificar a niños en riesgo de sufrir retrasos en el desarrollo y, por lo tanto, posee un alto nivel de sensibilidad para detectar niveles de desarrollo en el extremo inferior de la distribución. No obstante, nuestro propósito en este estudio es determinar si una prueba es apta para ser aplicada en la evaluación de intervenciones y, en consecuencia, analizar su capacidad para medir el desarrollo del niño a lo largo de toda la distribución de habilidades posibles, incluso en niños

con niveles de desarrollo altos y muy altos. En función de esto, modificamos la administración de esta prueba de la siguiente forma: si el niño lograba la puntuación máxima en una escala, evaluábamos adicionalmente los primeros tres ítems nuevos (es decir, no coincidentes) del cuestionario subsiguiente. Esto aumentó la variabilidad en las habilidades de desarrollo infantil medidas por la prueba y disminuyó el porcentaje de niños en el techo de un 10,5-15,5% a un 1,7-4,8%, según el dominio. Por otra parte, en lugar de que el cuidador respondiera el cuestionario por su cuenta, los ítems se completaron mediante entrevista. Esto se dispuso de este modo debido a que en algunas familias el nivel educativo era bajo. Asimismo, en los casos en que el cuidador no pudiera dar una respuesta a alguno de los ítems o si el fraseo del ítem denotaba la necesidad de evaluar directamente al niño para observar su desempeño, el encuestador se encargaba de administrarlo. En el manual del ASQ-3 se recomienda administrar los ítems directamente, en especial si se cuenta con el apoyo de (para) profesionales capacitados para la administración. En otros estudios llevados a cabo en países de ingresos bajos y medios, se han realizado adaptaciones del ASQ-3 similares a las que hemos mencionado (Fernald *et al*. 2012).

En el ASQ-3 se asigna una puntuación por escala y por cuestionario. Para las respuestas ‘sí’, ‘a veces’ o ‘no todavía’ se asigna 10, 5 o 0 puntos, respectivamente, y luego se calcula el total. Los ítems que no se completaron son reemplazados por el promedio de la escala (1,2% de los niños de la muestra). Sin embargo, si más de un ítem está incompleto en una misma escala, esta no se computa (0,3% de los casos).

Dados los cambios que se realizaron en el protocolo estándar de administración, el tiempo de aplicación de esta prueba aumentó a unos 20 minutos, en promedio, en contraste con los 10-15 minutos de duración que figuran en el sitio web de la editorial. El ASQ-3 está disponible en español y el paquete de materiales de la prueba (*Starter Kit*), que incluye cuestionarios y hojas de respuesta fotocopiables, un CD con cuestionarios en PDF para imprimir y una guía para el usuario en inglés, tiene un valor de USD 275. El paquete de manipulativos (*Materials Kit*) tiene un valor de USD 295 e incluye alrededor de 20 juguetes, libros y otros objetos manipulativos. Está diseñado para estimular la participación del niño durante la prueba y ayudar a que se lleve a cabo una evaluación efectiva. Estos materiales son necesarios para evaluar las habilidades del niño de forma directa, aun así no es obligatorio utilizar los materiales de este paquete, sino que se pueden reemplazar por otros materiales manipulativos que posean características similares. No obstante, es importante tener en cuenta que, cuando se mide el desarrollo a escala, en especial, para la evaluación de intervenciones, es fundamental estandarizar los protocolos de administración de modo que se garantice que las diferencias entre los niveles de desarrollo no estén relacionadas con la idiosincrasia del evaluador, ni en la administración ni en la puntuación de la prueba. En este sentido se recomienda que todos los evaluadores cuenten con un paquete de materiales estandarizado.

1.2.2.2. Prueba de Tamizaje del Desarrollo de Denver, segunda edición (Denver-II)

El Denver-II (Frankenburg *et al.* 1990; Frankenburg *et al.* 1992) es una prueba de tamizaje diseñada para ser utilizada por médicos o profesionales en la primera infancia con el fin de examinar el desarrollo de niños desde su nacimiento hasta los 6 años de edad. Está conformada por cuatro escalas que se administran y se puntúan de manera independiente—lenguaje, motricidad fina/adaptativa, motricidad gruesa y personal-social. Para la evaluación de la mayoría de los ítems que abarca la prueba (68%) se requiere que el evaluador observe el comportamiento y el desempeño del niño durante su ejecución, si

bien algunos ítems pueden ser respondidos por los padres, en particular en las escalas de desarrollo personal-social (76%) y de lenguaje (38%).

Los ítems por evaluar en cada niño se determinan trazando una línea (línea de edad) sobre la hoja de respuesta que marca el punto de inicio de la prueba. Para cada escala, los ítems están ordenados de acuerdo con el grado de dificultad (ascendente). Por cada ítem que el niño realiza correctamente debe escribirse 'pasó' como clave de su desempeño, si no logra ejecutar la tarea se coloca 'falló'. En los reportes del cuidador se puede colocar un 'no oportunidad' en los casos en que el cuidador no haya observado el desempeño del niño; asimismo, en los ítems por administración se puede colocar un 'rehusó' si el niño se niega a realizar la actividad. Los niños con al menos un 'rehusó' en alguno de los ítems del costado izquierdo de la línea de edad se consideran 'no evaluables' y la escala no recibe ninguna puntuación (0,5% de la muestra). Esta prueba clasifica el desarrollo del niño como 'normal' o 'sospechoso', según su desempeño en relación con el de los niños de la población de referencia. Sin embargo, para calcular la validez concurrente necesitamos, para cada escala, un puntaje continuo que podamos correlacionar con los valores de las escalas del Bayley-III. Para ello, y para cada dominio, construimos un puntaje 'bruto' de la siguiente forma: asignamos el valor 1 a 'pasó' y el valor 0 a 'falló' y sumamos todas las respuestas, incluyendo en el cómputo los ítems anteriores al punto de partida (que no fueron administrados y a los que se les asignó el valor de 1). A las claves 'no oportunidad' y 'rehusó' se asignó el valor 0.

El tiempo de administración registrado oscila entre los 15 y los 20 minutos; no obstante, en nuestro estudio el tiempo promedio fue de 27 minutos. Es probable que el tiempo de administración reportado por los autores esté basado en el desempeño de pediatras o profesionales. El set de materiales tuvo un costo de USD 200 cuando se compró para este estudio e incluía un manual técnico (en inglés), un manual de capacitación en inglés, las hojas de respuesta en español, un DVD con instrucciones para su administración y una bolsa pequeña con los objetos manipulativos necesarios para administrar la prueba (no incluía hojas en blanco). Cada hoja de respuesta adicional costó USD 0,45. En 2015 la editorial suspendió la comercialización de esta prueba, aun así los manuales y hojas de respuesta se pueden descargar desde su sitio web. En este sitio también se puede encontrar una foto que muestra todos los objetos manipulativos (juguetes y otros materiales) que se necesitan para administrar la prueba, si bien estos ya no están en venta.

1.2.2.3. Inventarios del Desarrollo de Habilidades Comunicativas de MacArthur-Bates, versión abreviada (SFI y SFII)

Los Inventarios I y II del Desarrollo de Habilidades Comunicativas MacArthur-Bates en su versión en español (S-CDIs, por sus siglas en inglés) (Jackson-Maldonado *et al.* 2003) son instrumentos de reporte reconocidos para la evaluación del desarrollo del lenguaje en niños y niñas de habla hispana de 8 a 18 meses y de 16 a 30 meses, respectivamente. Las versiones abreviadas de los S-CDIs, el SFI y SFII, se validaron en México y se crearon como una alternativa para uso como tamizaje o para evaluaciones que requirieran un instrumento sencillo de usar (Jackson-Maldonado, Marchman y Fernald 2012). Utilizamos la lista de vocabulario del SFI para evaluar el lenguaje receptivo y expresivo—es decir, la cantidad de palabras que el niño 'entiende' y las palabras que 'entiende y dice', respectivamente—en niños de 8 a 18 meses de edad; y la lista del SFII para evaluar el

lenguaje expresivo—esto es, la cantidad de palabras que el niño 'dice'—en niños de entre 19 y 30 meses.[7]

El puntaje bruto se calcula sumando las palabras que el niño 'entiende', 'entiende y dice' o 'dice', según la lista. En el inventario SFI la puntuación correspondiente a la comprensión de palabras debe ser siempre igual a la de producción de palabras o mayor. No se cuentan los ítems que están en blanco.

Para administrar la prueba, solo se necesitó la lista de vocabulario. Se pueden solicitar estas listas al Consejo Consultivo del Inventario del Desarrollo de Habilidades Comunicativas (CDI, por sus siglas en inglés) en la Universidad de Stanford y su costo se determina en cada caso, según el uso que se le dé a la prueba.[8] El set completo del S-CDIs con los manuales incluidos tiene un valor de USD 90 y cada hoja de respuesta adicional cuesta USD 1. Dado que estos cuestionarios se diseñaron y validaron en México, es probable que se necesite reemplazar algunas palabras por aquellas que sean de uso más frecuente en la región en donde se administre la prueba y de este modo garantizar un equivalente lingüístico y funcional de cada palabra—por ejemplo, en Colombia 'punta' es la palabra más común para referirse a un clavo. En el estudio, la administración de cada lista de vocabulario, por entrevista al cuidador, demoró alrededor de 8 minutos en promedio.

1.2.3. Pruebas cortas por validar en la Batería B

1.2.3.1. Inventario del Desarrollo de Battelle (versión de tamizaje), segunda edición (BDI-2)

La prueba de tamizaje BDI-2 (Newborg 2005) se creó con la finalidad de identificar posibles riesgos de rezago en el desarrollo de niños menores de 8 años. Este instrumento está compuesto por cinco escalas—cognitiva, comunicación, motora (que combina motricidad fina y gruesa), personal-social y habilidades adaptativas—que se administran y se califican de manera independiente. El procedimiento de aplicación recomendado para cada ítem se indica en la hoja de repuesta y puede ser de tres formas distintas: (i) *administración estructurada*, se evalúa la escala directamente en el niño; (ii) *observación* de las habilidades del niño por un período prolongado (por lo general durante la entrevista), y (iii) *entrevista* con el cuidador.[9]

Los ítems dentro de las escalas están organizados en orden ascendente de acuerdo con el grado de dificultad que presentan y existen criterios de inicio y fin (techo) para determinar el número de ítems sobre los que se evalúa al niño. Para cada ítem que el niño no logra completar se asigna el valor 0; si lo completa parcialmente se otorga 1 punto, y si lo completa en su totalidad se otorgan 2 puntos. El puntaje bruto es la suma de todas las respuestas correctas e incluye los ítems anteriores al nivel de base a los que se les asignan 2 puntos. Los ítems que no se hayan completado reciben el valor 0 (1,8% de los casos).

En el manual se recomienda que el evaluador tenga estudios universitarios, preferentemente en el área de la psicología o en disciplinas relacionadas. Sin embargo, también pueden administrar la prueba no profesionales que hayan recibido una capacitación rigurosa y supervisada sobre cómo administrar la prueba y cómo medir el desarrollo infantil. Según lo que establece el manual, el tiempo de aplicación varía entre 10 y 30 minutos. En

[7] Actualmente existe una versión del SF para niños de 30 a 37 meses, la cual se desarrolló en los meses entorno a la ejecución de nuestro estudio. Sin embargo, no supimos sobre esta versión sino luego de haber completado la recolección de datos.
[8] http://mb-cdi.stanford.edu/board.html.
[9] En los 31 ítems en los que se requería un período de observación prolongado (días o semanas), se sustituyó la 'observación' por la 'entrevista' como procedimiento de administración preferido.

nuestro estudio, el tiempo de administración se extendió considerablemente: en promedio, administrar la prueba completa tuvo una duración de 59 minutos por niño. Es probable que los tiempos de administración que se mencionan en el manual se hayan registrado sobre la base del desempeño de profesionales en distintas disciplinas relacionadas con esta área que tengan experiencia en la evaluación de niños y niñas. En todo caso, 10 minutos (el límite de tiempo más bajo registrado en el manual) es muy poco tiempo para administrar un promedio de nueve ítems en cinco escalas. En este estudio, observamos que el tiempo de aplicación aumentaba con la edad en los niños más pequeños y hasta los 24 meses de edad.

El set de materiales del BDI-2 cuesta USD 405,70 e incluye: el manual para el evaluador, los cuadernos de aplicación por área, un set de tarjetas, un cuaderno de estímulos, un paquete con 30 hojas de respuesta y los objetos manipulativos necesarios para administrar la prueba. Cada hoja de respuesta adicional tiene un valor de USD 3,08. Fue necesario adaptar y traducir parte del contenido de estos materiales, ya que la versión en español de la prueba contiene varias partes en inglés. Por ejemplo, hubo que traducir del inglés al español los cuadernos de aplicación (manuales), en los que se especifican las instrucciones para administrar los ítems y puntuarlos de manera correcta. Asimismo, también se tradujo el texto que aparecía en el libro de imágenes (libro de cuentos). Para poder comprar la prueba, el profesional debe mostrar a la editorial su certificación de estudios pertinente y acreditar su experiencia profesional en esta área.

1.2.3.2. Hitos del Desarrollo Motor Grueso de la Organización Mundial de la Salud (WHO-Motor)

El WHO-Motor (WHO Multicentre Growth Reference Study Group 2006; Wijnhoven *et al.* 2004) incluye seis hitos destinados a evaluar el desarrollo motor grueso en niños de 6 a 18 meses. No obstante, el análisis se realizó en niños de 6 a 15,9 meses, dado que el 91,9% de los niños más grandes logró alcanzar todos los hitos. La evaluación se llevó a cabo en forma directa y no se utilizaron reportes del cuidador para recolectar datos sobre la fecha (o la edad) en que los niños alcanzaron cada hito.

Debido a que esta prueba no ofrece indicaciones acerca de cómo calcular un puntaje bruto, sumamos la cantidad de hitos que el niño realizó e incluimos en el total los hitos anteriores. Por datos incompletos o inconsistencias, tres casos (1,4%) fueron descartados.

La versión en inglés del WHO-Motor se puede conseguir sin cargo en la página oficial de la OMS. Tradujimos al español las hojas de respuesta y las instrucciones de aplicación.

1.2.4. Nota breve sobre prematuridad

En ningún caso se ajustó por prematuridad antes de iniciar la evaluación. Es decir, en los niños prematuros se utilizó el mismo criterio para determinar el punto de partida que en el resto de los niños de la muestra, y luego de haber comenzado la prueba se retrocedía a los ítems anteriores (más fáciles) conforme con el nivel de desarrollo del niño (desempeño). A pesar de que esto aumenta el tiempo de administración, evita que el evaluador se base en el reporte del cuidador sobre la edad gestacional del niño para definir el punto de inicio de la prueba. Esto aporta mayor confiabilidad dado que frecuentemente los reportes sobre edad gestacional son erróneos.[10] La única excepción a esta regla fue el ASQ-3. Para esta prueba

[10] De hecho, se observó un 9% de inconsistencias (sobre el 50% de los niños que fueron reportados como prematuros) entre las semanas de gestación que se registraron en las encuestas y las que se registraron en el Bayley-III.

se siguieron los protocolos de aplicación del manual, que incluyen un ajuste por prematuridad, debido a que cada cuestionario se administra a un rango etario determinado.

1.2.5. Perfil y capacitación del evaluador y encuestador

Seis mujeres graduadas en psicología, algunas con experiencia previa en la evaluación de niños, recibieron capacitación sobre el Bayley-III durante seis semanas. Este período de capacitación incluyó entre veinte y veinticinco prácticas por evaluadora en niños cuyas edades estaban dentro del rango etario de la muestra de estudio. Ninguna de las seis evaluadoras conocía la prueba. Además, recibieron una capacitación de dos días y medio acerca de cómo medir la talla y peso de los participantes (y realizaron entre diez y doce prácticas). Ocho mujeres sin estudios universitarios ni experiencia previa en la evaluación de niños, recibieron capacitación sobre las pruebas cortas de la batería A o de la batería B. La capacitación duró entre seis y siete semanas, incluyendo capacitación en la encuestas de hogar. En promedio, cada encuestadora practicó la aplicación de cada prueba corta unas veinte veces (en la mayoría de los casos administraba toda la batería, A o B, en función de aquella en la que hubiera sido capacitada).

Las prácticas, tanto para las encuestadoras como para las evaluadoras, se llevaron a cabo en parejas y, en cada instancia, se evaluó la confiabilidad interobservador (nivel de concordancia) entre el capacitador y el aprendiz y entre las encuestadoras/evaluadoras de cada pareja. Para garantizar que las pruebas se administren de forma estandarizada, es conveniente continuar las prácticas hasta que la confiabilidad interobservador alcance un nivel satisfactorio (coeficiente de correlación intraclase, *CCI*>0,9) en cada escala de cada prueba. La Tabla 1 indica la cantidad de días de capacitación que se necesitan para cada prueba en promedio. Estas cifras se basan en la experiencia que hemos adquirido en este y en otros estudios y puede variar en función de la formación académica y experiencia previa de los encuestadores/evaluadores. La cantidad de prácticas depende también del grado de complejidad de la prueba y aumenta según el número de ítems que se deba evaluar en el niño o que se deba completar mediante observación directa, en comparación con aquellas pruebas que se administran mediante reporte del cuidador. Acorde con esto, los capacitadores coinciden en reportar que resulta más fácil enseñar y administrar aquellas pruebas y escalas en las que la mayoría de los ítems se responden mediante reporte del cuidador. Se trabajó con tres capacitadores, uno por cada serie de pruebas: el Bayley-III, las pruebas cortas de la batería A y las pruebas cortas de la batería B. Los tres capacitadores cuentan con una maestría en psicología.

Durante el proceso de recolección de datos, el 5% de las administraciones fueron observadas y evaluadas por el capacitador correspondiente y se calculó la confiabilidad interobservador. Luego de observar estas administraciones, el capacitador realizaba los comentarios y correcciones correspondientes respecto del desempeño de la encuestadora/evaluadora. La concordancia entre las puntuaciones de la encuestadora/evaluadora y el capacitador durante estas pruebas fue alta (*CCI* promedio =0,95), lo que indica que la calidad en la administración se mantuvo a lo largo de toda la recolección de datos.

1.3. Análisis de la muestra

En la Figura 1 se puede observar el flujo de participantes en el estudio y el número de sujetos que intervinieron en cada evaluación. Los datos se recogieron sobre una muestra de 1.533 niños y niñas de 6 a 42 meses en 497 manzanas, en su mayoría dentro de los

estratos 1-3 de Bogotá.[11] No obstante, el Bayley-III se administró en 1.330 niños (86,8%), para quienes también contamos con la encuesta de hogar y su resultado en las pruebas cortas. El 13,2% de los niños restantes, a quienes no se les pudo administrar el Bayley-III, tienen una mayor probabilidad de asistir a un centro de cuidado infantil, tener una madre joven o vivir en hogares con niños mayores que ellos y sin personas de la tercera edad. Esto sugiere que para aquellas madres que no cuentan con formas alternativas de cuidado para sus (otros) hijos pudo haber sido difícil encontrar el tiempo para llevar al niño a que le realicen la prueba. Rubio-Codina *et al.* (2015) señala que la muestra resultante de niños a quienes se les administró el Bayley-III continuó siendo representativa por estrato socioeconómico de los hogares.

De los 1.330 niños que fueron evaluados mediante el Bayley-III, 4 (0,3%) no completaron la prueba y 15 (1,1%) obtuvieron puntuaciones <70 en alguna de las escalas de la prueba y por esto fueron excluidos. Los 1.311 niños restantes que aportaron datos completos y consistentes respecto del Bayley-III constituyen nuestra muestra de estudio. A estos se les administraron las pruebas cortas correspondientes a cada edad de la batería A (n_A =676) o batería B (n_B =635), sobre la base de una asignación aleatoria luego de haber estratificado la muestra por edad y sector socioeconómico.

La Tabla 2 presenta información estadística de un conjunto de características del niño, de sus padres y del hogar organizadas por batería. La última columna indica para cada variable el valor *p* de la diferencia del promedio entre las dos baterías. Alrededor del 15% de los niños en la muestra son prematuros (edad gestacional <37 semanas) y el 17-18% presentan baja talla por edad (puntaje z de talla por edad <-2 desviaciones estándar (DE) en relación con la mediana del puntaje z establecida por los Estándares de Crecimiento de la OMS (WHO 2006)).[12] En promedio, las madres tienen entre 26 y 27 años de edad y poseen un nivel de escolaridad de un poco más de diez años. El 30-31% de ellas superó el nivel secundario y el 50-54% reportaron tener empleo, ya sea remunerado o no remunerado. El nivel educativo de los padres es menor que el de las madres y en un 32-34% de los casos los padres no vivían en el hogar con sus hijos.

La distribución de los participantes según edad y sexo (Panel I) y estratos socioeconómicos (Panel III) en ambas baterías está balanceada, conforme a lo esperado dado el diseño del estudio, es decir, conforme a la asignación aleatoria por grupo etario y estrato socioeconómico. A pesar de que otras características del niño y de los padres también se encuentren distribuidas de manera balanceada entre las dos baterías, los hogares en la batería B presentan un índice de riqueza de hogar ligeramente más elevado que los de la batería A (*p* =0,013, Panel III). Esto es acorde con la diferencia estadísticamente significativa de 0,6 años de educación adicionales que reportaron los padres de la batería B con respecto a los padres de la batería A (*p* =0,026, Panel II). Es posible que estas diferencias estén relacionadas con la gran heterogeneidad en el nivel de riqueza del hogar existente dentro de cada estrato, documentada en Rubio-Codina *et al*. (2015). El índice de riqueza del hogar se obtiene mediante el análisis de componentes principales con correlaciones policóricas sobre una serie de activos del hogar y características de la vivienda, tal y como se hizo en Rubio-Codina *et al.* (2015).[13] La calidad del entorno del

[11] Se evaluaron a 403 niños en 134 manzanas del Estrato 1, a 459 niños en 159 manzanas del Estrato 2, a 457 niños en 199 manzanas del Estrato 3 y a 12 niños en cinco manzanas del Estrato 4.
[12] Se ajustaron todas las mediciones a los estándares de crecimiento de la OMS utilizando el software WHO Anthro (versión 3.2.2, 2011).
[13] Entre las variables se encuentran: auto, refrigerador, microondas, lavadora, calentador, computador, teléfono inteligente, televisor de pantalla plana, sistema de sonido envolvente, reproductor de DVD, estéreo, consola de juegos, internet, garaje, cocina compartida con otras familias, baño compartido, posee más de un baño, posee pisos de calidad (baldosa, alfombra o pisos de madera en comparación con pisos de grava, cemento o tierra), posee ventanas exteriores y el índice de cohabitación.

hogar, medida mediante los indicadores de variedades de actividades y materiales de juego del FCI también es similar entre los hogares de ambas baterías.[14] Es importante mencionar que los niños de las baterías A y B presentan niveles de desarrollo semejantes, según lo que indican los resultados del Bayley-III (Panel IV).

2. Análisis estadístico

2.1. Estandarización interna de las puntuaciones

Para cada escala construimos un puntaje bruto siguiendo las instrucciones de los manuales de administración de las pruebas y los pasos indicados previamente. Debido a que el Denver-II no cuenta con un puntaje bruto, se sumaron los ítems que el niño realizó con éxito y los ítems anteriores al nivel de base, de acuerdo con los principios generales de puntuación. Bajo este mismo criterio, a fin de establecer un puntaje bruto en la escala del WHO-Motor, se sumaron todos los hitos que el niño logró, incluyendo los hitos anteriores. Por consiguiente, dada su composición, el puntaje bruto incrementa con la edad en todas las pruebas, a excepción del ASQ-3 cuyos cuestionarios son edad-específicos. La Tabla A1 del Apéndice III indica el puntaje bruto de las pruebas cortas para todos los niños de 6 a 42 meses, dividiendo este rango etario en intervalos de doce meses—esto es, niños de 6 a 18 meses, de 19 a 30 meses y de 31 a 42 meses. En los valores expuestos en la tabla se evidencia el gradiente de edad del puntaje bruto promedio. En el caso del ASQ-3, las puntuaciones oscilan de forma arbitraria y tienden a incrementar con la edad, en especial en el grupo de niños más grandes.

A fin de controlar por los efectos de edad, es necesario estandarizar por edad el puntaje bruto. No obstante, ninguna de las pruebas cortas ni tampoco el Bayley-III habían sido estandarizados (normados) para Colombia. Por otra parte, en esta muestra, las puntuaciones compuestas del Bayley-III, estandarizadas externamente, varían según la edad de manera inusual, disminuyendo en la escala cognitiva, aumentando en la escala de motricidad y adoptando una distribución en forma de U en la escala del lenguaje (Rubio-Codina *et al.* 2015). Adicionalmente, las desviaciones estándar son inferiores a los 15 puntos esperados en la población estandarizada y disminuyen con la edad, en particular en la escala cognitiva. Estos patrones, que también se pueden observar en la Tabla A1 del Apéndice III, sugieren que las normas externas del Bayley-III (desarrolladas sobre la base de una muestra representativa de la población estadounidense) no son las adecuadas para esta muestra. Se puede llegar a conclusiones similares respecto de la fluctuación en los promedios del ASQ-3 y en las desviaciones estándar descritas recientemente. Por consiguiente, tal y como es habitual en el trabajo con datos de países en desarrollo, se estandarizaron internamente las puntuaciones por edad. Estandarizar internamente las puntuaciones procediendo del mismo modo para todas las pruebas permite controlar los efectos relativos a la edad de manera consistente (para todas las pruebas), lo que facilita el poder establecer comparaciones entre las pruebas. Aplicar las normas de las poblaciones de referencia para cada prueba (estandarización externa) no concede esta ventaja.

El primer componente principal explica el 43,09% de la varianza total y el segundo componente principal explica el 8,03% de la varianza adicional.

[14] La variedad en los materiales de juego se calcula a través de la suma de los siguientes indicadores: juguetes para crear o reproducir música; objetos para apilar o construir; materiales para dibujar, escribir, colorear o pintar; juguetes para moverse y desplazarse; juguetes para el juego de roles; libros para colorear y dibujar y libros de imágenes para niños, y juguetes para aprender formas y colores. La variedad en las actividades de juego se calcula a través de la suma de los siguientes indicadores: leer libros u observar las ilustraciones en ellos con el niño; contarles cuentos a los niños; cantar canciones con el niño; jugar con el niño utilizando sus juguetes; pasar tiempo con el niño haciendo garabatos, dibujando o coloreando; pasar tiempo con el niño nombrando o contando objetos, y salir a pasear con el niño.

Por lo general, para llevar a cabo la estandarización interna se divide la muestra en grupos etarios lo más pequeños posible—de forma mensual, preferentemente, dado el alto nivel de sensibilidad que demuestran los hitos del desarrollo respecto de la edad durante la etapa temprana del niño. Al mismo tiempo, se deben garantizar que haya suficientes observaciones por grupo y calcular el puntaje z dentro de cada grupo etario (véase Fernald *et al.* 2011, por ejemplo). En este estudio, adoptamos este enfoque, sin embargo los puntajes z internos se estimaron con mayor flexibilidad, teniendo en cuenta el tamaño limitado de la muestra. Concretamente, en lugar de utilizar promedios y desviaciones estándar específicos para cada mes de edad, se estimaron los promedios y desviaciones estándar condicionados a la edad mediante métodos no paramétricos, tal como se describe en el Apéndice II. Este método demuestra una menor sensibilidad a observaciones extremas y a tamaños de muestra pequeños dentro de cada grupo etario, y a su vez permite reproducir con mayor precisión el modo en que estas pruebas calcularían los puntajes externos (dado su carácter completamente no paramétrico).[15] Nótese que, a fin de corregir los efectos relacionados con la idiosincrasia del evaluador/encuestador sobre la administración y puntación de las pruebas, estandarizamos internamente los *residuos* de los puntajes brutos por edad, netos del efecto de evaluadores/encuestadores, en lugar de estandarizar de forma directa los puntajes brutos.

2.2. Análisis de la confiabilidad, validez y viabilidad de las pruebas

Luego de proporcionar evidencia empírica que respalda la validez del Bayley-III como nuestro patrón de referencia en este estudio, pasamos al análisis de la confiabilidad, validez y viabilidad de las pruebas cortas.

En primera instancia, analizamos por escala la confiabilidad del test-retest y de la consistencia interna de las pruebas cortas. La confiabilidad es el grado en que un instrumento de evaluación produce resultados estables y consistentes. La confiabilidad del test-retest permite medir el grado de estabilidad de una prueba a través del tiempo. Esta medida se obtiene calculando el coeficiente de correlación intraclase entre las puntuaciones de una escala que fueron recogidas en dos instancias de evaluación diferentes, administradas en el mismo niño y por el mismo evaluador/encuestador, con una diferencia de unos pocos días entre una y otra, por lo general en un lapso de una a dos semanas. La confiabilidad de la consistencia interna examina hasta qué punto los ítems de una escala (o prueba) miden el mismo constructo subyacente (dominio o habilidad). Esta medida de confiabilidad se puede estimar calculando el coeficiente alfa de Cronbach (□) en todos los ítems de la escala para todos los niños en la muestra (de 6 a 42 meses) y en grupos etarios de intervalos de doce meses.[16] Los índices más altos de confiabilidad indican un mejor desempeño de la prueba en una población determinada. El análisis de confiabilidad reviste especial importancia cuando una prueba se administra en una población para la cual no fue diseñada, en especial si el idioma o el contenido de los ítems fueron modificados a fin de asegurar su comprensión y equivalente lingüístico. Asimismo, estudiamos en qué medida las escalas de una prueba se correlacionan entre sí, lo cual, a su vez, es un indicador del grado de congruencia que se establece entre las escalas y de la interrelación que existe entre los distintos dominios del desarrollo.

[15] Por ejemplo, en el caso del Bayley-III, las puntuaciones compuestas son una función no lineal de los puntajes brutos. En particular, (i) se da un mayor peso a los ítems administrados en niños de edades más tempranas, y (ii) se establecen normas por grupos de un mes de edad hasta los 36 meses y a partir de ahí en intervalos más largos.

[16] Dado que los cuestionarios del ASQ-3 son edad-específicos, los ítems varían de cuestionario en cuestionario. Por lo tanto, se estimó la consistencia interna de cada cuestionario por separado y se promedió el □ para todos los cuestionarios en el grupo etario de interés.

El siguiente paso es el análisis de validez. Esta hace referencia al desempeño que muestra una prueba para medir lo que se supone debe poder medir. En general, se considera el elemento de mayor importancia en pruebas psicológicas, dado que afecta la relevancia que se asigna a los resultados obtenidos a partir de las pruebas y su interpretación. En este estudio, nos centraremos en la validez de criterio—es decir, en la correlación de los resultados de las pruebas con otro criterio de interés. La validez de criterio puede ser concurrente o predictiva, según se refiera al desempeño actual de la prueba o a una predicción futura de este. Dado que los datos utilizados en este estudio son de sección cruzada, solo se puede estudiar la validez de criterio concurrente, a la que de ahora en adelante denominaremos 'validez concurrente'.

Comenzamos analizando la validez concurrente entre las puntuaciones en cada escala de cada prueba y un conjunto de variables teóricamente relacionadas con el desarrollo infantil mediante el cálculo del coeficiente de la correlación de Pearson (*r*) por dominio (escala). Dentro del conjunto de variables consideradas se incluyen la educación de la madre, el índice de riqueza del hogar, las puntuaciones FCI de las actividades y materiales de juego en el hogar y dos indicadores para prematuridad y baja talla por edad, respectivamente. Luego nos centramos en la base del estudio, es decir, en el análisis de la validez concurrente entre las pruebas cortas y la escala de Bayley-III (nuestro patrón de referencia). Esto se lleva a cabo calculando la correlación de Pearson (*r*) por dominio y por grupo etario. Debido a que todas las correlaciones utilizan puntajes estandarizados internamente—*i. e.* estandarizados internamente por edad una vez eliminados los efectos del evaluador/encuestador—esto equivale a calcular las correlaciones parciales controlando por los efectos del evaluador/encuestador y de la edad de manera flexible. Los valores *p* para las correlaciones se calculan utilizando métodos de *bootstrap* (de remuestreo) con 1.000 repeticiones y definiendo conglomerados (*clusters*) por edad y sector (Efron 1982). Siguiendo a Evans (1996), clasificamos las correlaciones de Pearson como bajas (*r*=0,20-0,39), moderadas (*r*=0,40-0,59) y altas (*r*=0,60-0,79) a lo largo de la presentación de los resultados y de la discusión.

También utilizamos los métodos *bootstrap* para comparar el tamaño de las correlaciones de cada prueba corta con el Bayley-III por grupo etario y para identificar aquellas que son estadísticamente distintas entre sí. Por ejemplo, evaluamos si la correlación entre las escalas cognitivas del BDI-2 y las del Bayley-III son significantemente mayores o menores que la correlación entre las escalas cognitivas del ASQ-3 y las del Bayley-III. Esto se lleva a cabo para cada par de correlaciones (dentro de un mismo grupo etario) que presenta una diferencia lo suficientemente mayor como para ameritar un análisis de significancia estadística.

Asimismo, realizamos una variedad de chequeos de robustez. Empezamos analizando la robustez de los resultados obtenidos en el análisis de la validez concurrente al uso de las puntuaciones compuestas del Bayley-III (estandarización externa) y al uso de métodos paramétricos para la estandarización. Luego, repetimos este análisis utilizando la versión original de seis ítems del ASQ-3, controlando por los efectos de prematuridad antes de realizar la estandarización de las puntuaciones, y dividiendo la muestra en grupos etarios de intervalos de 6 meses de edad. Por último, calculamos las correlaciones de Spearman (rango) a fin de estudiar la monotonicidad en la relación entre dos puntuaciones cualesquiera, en contraposición con la correlación lineal, y las correlaciones canónicas (es decir, las correlaciones entre conjuntos de variables—en este caso, las escalas en una prueba) para controlar por el carácter multidimensional de las medidas analizadas. Esto también corrige por el gran número de correlaciones examinadas. Debido a limitaciones de

espacio, incluimos algunas de estas pruebas en el Apéndice III y otras se encuentran a disposición bajo solicitud.

Posteriormente, investigamos si la concurrencia entre las pruebas cortas y el Bayley-III varía según el estatus socioeconómico del hogar. Repetimos el análisis de la validez concurrente por dominio para los hogares situados dentro del 25% más bajo y del 25% más alto de la distribución de riqueza por separado. Como se especificó anteriormente, evaluamos mediante métodos *bootstrap* si el tamaño de la correlación entre el 25% de los hogares más pobres de la muestra y el 25% de los más ricos presenta una diferencia estadísticamente significativa. Nos centramos en la validez concurrente para los dominios del desarrollo coincidentes únicamente. Asimismo, trabajamos con la totalidad de la muestra (niños de 6 a 42 meses) dado su tamaño limitado y con el fin de evitar rechazar la hipótesis nula de no diferencia en las correlaciones por estatus socioeconómico por falta de poder estadístico.

Por último, analizamos la viabilidad de la administración de las pruebas. Esto comprende todos los costos relacionados con la compra y administración de cada una de ellas, incluidos aquellos gastos que surgen durante su adaptación y capacitación.

3. Resultados

3.1. El Bayley-III como patrón de referencia

La Tabla A1 del Apéndice III muestra que los promedios de las puntuaciones compuestas del Bayley-III se encuentran dentro de sus valores normales, a pesar de que exhiben una relación inusual con la edad. Las desviaciones estándar son más bajas de lo esperado y además disminuyen con la edad, en particular para la escala cognitiva. Como se indica en la sección anterior, estos datos confirman una vez más la importancia de llevar a cabo una estandarización interna.

Los primeras dos filas en la Tabla 3 muestran las confiabilidades de test-retest y de consistencia interna de los puntajes brutos y puntuaciones compuestas para las escalas del Bayley-III. El grado de confiabilidad del test-retest en veinte niños luego de entre seis y diecinueve días (un promedio de ocho días) es alto, todos los *CCI* ≥0,96, lo que indica una muy buena estabilidad temporal del Bayley-III. De modo similar, la consistencia interna parece ser buena por rango etario para todos los dominios (escalas), incluso si no se observa un patrón claro respecto de la edad. Ambas medidas de confiabilidad son más elevadas para el caso del Bayley-III que para cualquiera de las pruebas cortas, lo que respalda el hecho de que haya sido elegido como patrón de referencia.

Además, las correlaciones entre las escalas del Bayley-III para la totalidad de la muestra de niños de 6 a 42 meses (primera fila, Tabla 5) y por grupo etario (primera fila, Tabla 6) son similares a aquellas que se reportan en el manual de la prueba (Bayley 2006). Estas exhiben una leve tendencia a aumentar con la edad.

3.2. Confiabilidad de las pruebas cortas

Las filas de la tres a la nueve en la Tabla 3 muestran las confiabilidades de test-retest y de consistencia interna para las pruebas cortas. La consistencia interna se reporta primero para toda la muestra y luego por grupo etario.

El grado de confiabilidad del test-retest se obtuvo luego de dos a once días (un promedio de ocho días) para doce niños en las pruebas cortas de la batería A; y luego de cinco a once días (un promedio de siete días) para once niños en las pruebas de la batería B. A pesar de que los tamaños de la muestra son reducidos, los valores son en general satisfactorios (CCI

≥0,7) e indican mediciones estables a través del tiempo para la mayoría de las pruebas. Las únicas excepciones son las escalas de motricidad fina en las versiones de seis y nueve ítems del ASQ-3 (ambas *r* =0,37) y las escalas de motricidad gruesa (*r* =0,53) y personal-social (*r* =0,49) del Denver-II.

Si se considera a todos los niños de la muestra, los coeficientes alfa de Cronbach (□) son generalmente altos y por encima del punto de corte deseado de 0,7, lo que sugiere niveles de consistencia interna buenos□□Las únicas excepciones son las versiones de nueve y seis ítems del ASQ-3, que obtuvieron valores inferiores a este corte en todas las escalas, salvo en la escala de motricidad gruesa de la versión de nueve ítems. Tal y como se observó en los valores del Bayley-III, la consistencia interna de las pruebas cortas no presenta ningún patrón consistente respecto de la edad, si bien los valores en general son más bajos entre los grupos etarios que cuando se considera la totalidad de la muestra. Por grupo etario, la mayoría de las escalas del ASQ-3 y del BDI-2 muestran una consistencia interna baja o muy baja, en especial para los niños del grupo etario intermedio en el caso del BDI-2. Investigamos también si había algún ítem específico que fuera particularmente problemático, es decir, algún ítem que no mida el mismo constructo subyacente y que pudiera explicar la baja consistencia interna de algunas pruebas; pero no se logró identificar nada al respecto.

Nótese que la consistencia interna de la versión de seis ítems del ASQ-3 es considerablemente inferior a la consistencia interna de la versión de nueve ítems para cada escala y grupo etario. De hecho, todas las escalas excepto una exhiben una consistencia interna baja, con valores de ≤0,55. Esto respalda nuestra elección de utilizar la versión de nueve ítems y, por tal motivo, todos los resultados que se reporten de aquí en adelante serán en función de esta.[17]

Las correlaciones entre las escalas del ASQ-3, Denver-II y BDI-2 (disponibles bajo solicitud) son por lo normal inferiores a las correlaciones observadas entre las escalas del Bayley-III, ya sea en la totalidad de la muestra o por grupo etario, lo que justifica una vez más la preferencia del Bayley-III como patrón de referencia de este estudio. En el caso del ASQ-3, los valores también son un poco más bajos que los que se reportan en el manual. Esto puede deberse a las adaptaciones y modificaciones que se efectuaron en la administración de la prueba. Es importante tener en cuenta también que esta comparación es parcial dado que las correlaciones reportadas en el manual abarcan de 1 a 66 meses de edad, mientras que aquellas que se pueden calcular con los datos disponibles para este estudio abarcan un rango etario limitado de niños más pequeños y, por lo general, la interrelación entre las escalas tiende a aumentar con la edad. No se puede realizar esta comparación para el Denver-II y el BDI-2 debido a que estas correlaciones no figuran en sus respectivos manuales.

3.3. Correlaciones con otras variables

En un primer análisis de la validez concurrente, calculamos la correlación de cada escala de las pruebas con un conjunto de variables socioeconómicas que se supone que están relacionadas con el desarrollo del niño y que se ha demostrado empíricamente que en efecto están correlacionadas con estas variables en una variedad de contextos y países, incluida la región de América Latina. Entre las variables se consideran las siguientes:

[17] Los resultados obtenidos utilizando la versión de seis ítems son robustos a los de la versión de nueve ítems y están a disposición bajo solicitud.

educación de la madre, índice de riqueza del hogar, calidad del entorno del hogar—medido a través de las puntuaciones FCI para materiales y actividades de juego disponibles en el hogar—prematuridad y baja talla por edad.

La primera fila en la Tabla 4 muestra que todas las escalas del Bayley-III se correlacionan significativamente con las primeras cuatro variables, a excepción de la escala de motricidad gruesa que presenta correlaciones muy bajas y significativas con la educación de la madre, materiales de juego y prematuridad únicamente. Las correlaciones con prematuridad y baja talla por edad son bajas en todas las escalas y no siempre son significativas. Más concretamente, las correlaciones entre la escala de motricidad fina y prematuridad y entre las escalas cognitiva y de motricidad gruesa y baja talla por edad no son significativas. La existencia de estas correlaciones bajas puede deberse a que en la muestra el número de niños prematuros o con baja talla por edad es relativamente bajo.

Las filas subsiguientes presentan las correlaciones para las pruebas cortas. Como se ha demostrado, la mayoría de las escalas presentan correlaciones bajas pero significativas con al menos dos de los factores socioeconómicos analizados, a excepción de todas las escalas de motricidad gruesa, así como también de la escala de lenguaje expresivo del SFI en niños menores de 18 meses y la escala personal-social del Denver-II. Las correlaciones entre las escalas de las pruebas cortas y las variables prematuridad y baja talla por edad por lo general no son significativas, salvo para las escalas de comunicación, motricidad y habilidades adaptativas del BDI-2. En términos generales, el BDI-2 exhibe los valores de correlación más elevados entre las pruebas cortas. No obstante, estos tienden a ser más bajos que los valores observados para el Bayley-III.

El análisis de correlaciones por grupo etario (disponible bajo solicitud) muestra que por lo general las correlaciones tienden a aumentar con la edad, en particular aquellas que se establecen entre el desarrollo infantil y las variables educación de la madre y riqueza del hogar. Esto es consecuente con los gradientes socioeconómicos del desarrollo infantil reportados por numerosos autores en la región y a nivel mundial (Schady *et al.* 2015; Fernald *et al.* 2011; Hamadani *et al.* 2014; Rubio-Codina *et al.* 2015).

3.4. Validez concurrente

3.4.1. Escalas de un mismo dominio, totalidad de la muestra

Las correlaciones promedio entre las escalas de las pruebas cortas y el Bayley-III para todo el rango etario de la muestra (de 6 a 42 meses) se exhiben en la Tabla 5. Aquellas correlaciones entre las escalas que miden el mismo dominio del desarrollo están resaltadas en negrita. Dado que el Denver-II no contiene una escala cognitiva, correlacionamos la escala de motricidad fina-adaptativa con la escala cognitiva del Bayley-III. Además, a excepción del SFI, el resto de las pruebas cortas incluye dentro de una misma escala (escala de comunicación/lenguaje) los ítems relacionados con el lenguaje receptivo y expresivo. Correlacionamos esta escala con ambas escalas de lenguaje del Bayley-III (lenguaje receptivo y expresivo). Del mismo modo, la escala de motricidad del BDI-2 combina los ítems de motricidad fina y gruesa y, por lo tanto, correlacionamos esta escala con ambas escalas de motricidad del Bayley-III (motricidad fina y gruesa). Asimismo, como se indicó previamente, no existe una escala del Bayley-III que se corresponda con la escala personal-social o adaptativa de las pruebas cortas y permita llevar a cabo un análisis.

En general, los resultados muestran que la magnitud de las correlaciones entre el Bayley-III y las pruebas cortas es de baja a moderada. Por dominio del desarrollo, dentro de las pruebas multidimensionales, las correlaciones más altas se observan en la escala de

lenguaje expresivo, seguida de la escala de motricidad gruesa. Las correlaciones en la escala de lenguaje expresivo son moderadas: *r* =0,506 entre la escala de lenguaje expresivo del Bayley-III y la subescala de lenguaje del Denver-II, *r* =0,495 con la escala de comunicación del BDI-2, y *r* =0,395 con la escala de comunicación del ASQ-3. Para la escala de lenguaje receptivo, las correlaciones siguen el mismo patrón que se observa en la escala del lenguaje expresivo, pero por lo general son entre un 20% y un 40% más bajas. En cuanto a las pruebas que miden un solo dominio del desarrollo, la concurrencia en lenguaje también es mayor con las escalas de lenguaje expresivo que con las de lenguaje receptivo. Esto es así para las escalas del SFII, pero no para las del SFI. Respecto del desarrollo motor grueso, las correlaciones entre la escala de motricidad gruesa del Bayley-III y su respectiva escala en las pruebas cortas son particularmente altas para el WHO-Motor (*r* =0,703), moderadas para el Denver-II (*r* =0,499) y bajas para el BDI-2 (*r* =0,339) y el ASQ-3 (*r* =0,325). Si bien es estadísticamente significativa, la concurrencia promedio para el desarrollo cognitivo y de motricidad fina entre las escalas correspondientes es por lo general baja para todas las pruebas cortas.

3.4.2. Escalas de un mismo dominio, por grupos etarios

La Tabla 6 muestra la validez concurrente por grupos etarios de intervalos de doce meses. Las letras (a, b, c, etc.) que aparecen junto a algunos de los valores de correlación indican si la correlación en cuestión es significativamente mayor (en términos estadísticos) que la correlación entre otra prueba corta (señalada en la nota al pie de la tabla) y el Bayley-III. Tal y como se explicó en la Sección 3.2, esto se lleva a cabo para cada par de correlaciones (dentro de un mismo grupo etario) que presenta una diferencia lo suficientemente grande como para ameritar un análisis de significancia estadística. El Gráfico 2 complementa este análisis, ya que representa por edades la correlación entre las escalas del Bayley-III y las escalas de las pruebas cortas que miden el mismo dominio del desarrollo. Dados los patrones de validez concurrente por edad que se observan en estos gráficos, a continuación se analizan los dominios del desarrollo cognitivo, de lenguaje y de motricidad fina separados del dominio de motricidad gruesa.

Cognitivo, lenguaje y motricidad fina. Como se puede observar en la Tabla 6, las escalas cognitiva/motricidad fina-adaptativa, de lenguaje/comunicación y de motricidad fina del Denver-II y del BDI-2 presentan correlaciones similares entre sí y bajas pero estadísticamente significativas con las correspondientes escalas del Bayley-III en niños de 6 a 18 meses (*r* =[0,164, 0,315]). En niños de 19 a 30 meses, la validez concurrente aumenta en poca medida (*r* =[0,256, 0,610]) y solo alcanza valores de moderados o altos en las escalas de lenguaje. La concurrencia continúa aumentando a partir de los 30 meses en todos los dominios (*r* =[0,380, 0,702]) y nuevamente se alcanzan las correlaciones más altas en la dimensión de lenguaje. En cuanto a este dominio, las correlaciones con la escala de lenguaje expresivo del Bayley-III siempre son más altas que con la escala de lenguaje receptivo en todo el rango etario.

Al comparar los valores de las correlaciones de las pruebas cortas multidimensionales con el Bayley-III, se puede observar que las escalas del ASQ-3 presentan de modo consistente correlaciones más bajas que las escalas del Denver-II y el BDI-2. Estas correlaciones son significativamente inferiores en 16 comparaciones de las 24 que constituyen el total (*P* <0,05). La única excepción es la correlación entre la escala de motricidad fina del ASQ-3 y la del Bayley-III en niños de 31 a 42 meses, que tiene la misma magnitud que la correlación con la escala de motricidad del BDI-2. En el grupo de niños más pequeños, las correlaciones del ASQ-3 son por lo general triviales y no son significativas en todos los

dominios. Además, la escala de resolución de problemas del ASQ-3 no predice significativamente la escala cognitiva del Bayley-III sino hasta los 31 meses. Las correlaciones con valores más elevados se pueden observar en la escala de comunicación del ASQ-3: estos son de bajos a moderados con la escala de lenguaje receptivo del Bayley-III (r =[0,231, 0,402]) y de moderados a altos con la escala de lenguaje expresivo (r =[0,458, 0,560]) en niños de 19 meses en adelante.

La escala de lenguaje expresivo del SFI presenta correlaciones un poco más elevadas con ambas escalas del lenguaje del Bayley-III en comparación con las correlaciones de la escala de lenguaje receptivo del SFI; sin embargo, estas diferencias no son estadísticamente significativas. Es posible que sea más fácil para las madres reportar las palabras que el niño dijo que aquellas que el niño comprendió. En el grupo de niños más pequeños, la escala de lenguaje expresivo del SFI muestra una correlación baja con la escala de lenguaje receptivo del Bayley-III (r =0,373). No obstante, esta correlación es significativamente mayor que aquella con las escalas de lenguaje/comunicación del Denver-II y el BDI-2 (ambas P <0,05) y con la escala de comunicación del ASQ-3 (P <0,001). La correlación entre la escala de lenguaje expresivo del SFI y su correspondiente escala en el Bayley-III (r =0,242) es similar a las correlaciones que se establecen con el resto de las pruebas cortas. En el intervalo de 19 a 30 meses de edad, la escala de lenguaje expresivo del SFII obtiene una correlación baja con la escala de lenguaje receptivo del Bayley-III (r =0,241). Este es un valor significativamente más bajo que el obtenido para el Denver-II (P <0,05). Sin embargo, la escala de lenguaje expresivo del SFII presenta una correlación alta con la misma escala del Bayley-III (r =0,600), similar a la del Denver-II y el BDI-2, y significativamente mayor que la del ASQ-3 (P <0,05).

Motricidad gruesa. Las escalas de motricidad gruesa presentan un comportamiento distinto del resto de los dominios del desarrollo. Como se puede observar en el Gráfico 2 y en la Tabla 6, la escala de motricidad del BDI-2 muestra correlaciones bajas con la escala de motricidad gruesa del Bayley-III, que varían muy poco de un rango etario a otro (r =[0,311, 0,371]). No obstante, las correlaciones entre las escalas de motricidad gruesa del Denver-II y el ASQ-3 son de moderadas a altas en niños de 6 a 18 meses (r =[0,585, 0,654]. Estos valores son significativamente mayores que los del BDI-2 (P <0,05) y disminuyen a partir de esa edad. Mientras que la concurrencia para el Denver-II disminuye a niveles moderados (r =[0,406, 0,426]), en el caso del ASQ-3, desciende a niveles bajos (r =[0,175, 0,218]), significativamente inferiores a los obtenidos para el Denver-II (P <0,05). Las correlaciones más bajas para la escala de motricidad gruesa del BDI-2 con su correspondiente escala en el Bayley-III, incluso en los niños más pequeños de la muestra, puede deberse a que la escala de motricidad del BDI-2 combina ambas habilidades motoras (fina y gruesa).

En los niños de 6 a 15 meses, el WHO-Motor posee una correlación alta con la escala de motricidad gruesa del Bayley-III (r =0,703). Esta correlación es más alta que en cualquier otra prueba para el dominio de motricidad gruesa, sin embargo, solo es significativamente mayor que la del BDI-2 (P <0,001).

3.4.3. Distintas escalas de dominios

En ocasiones, las correlaciones entre el Bayley-III y las pruebas cortas son más elevadas entre las escalas que miden distintas funciones que entre aquellas que miden las mismas funciones (Tabla 6). Esto sucede con menor frecuencia a medida que aumenta la edad del niño. En el grupo de los más pequeños, las escalas de desarrollo personal-social del

Denver-II y del ASQ-3 se correlacionan con las escalas cognitiva, de lenguaje y de motricidad fina. A menudo, en los niños de más de 18 meses, las escalas de lenguaje se relacionan significativamente con la escala cognitiva del Bayley-III. En los niños de más de 30 meses, la correlación para la escala de lenguaje del Denver-II con la escala cognitiva es significativamente más alta que con la escala de motricidad fina-adaptativa (P <0,05). No se observan muchos más patrones claros en las intercorrelaciones.

Las pruebas que miden dominios del desarrollo específicos también se correlacionan con otros dominios. En el grupo de los más pequeños, la escala de lenguaje expresivo del SFI muestra correlaciones significativas pero bajas con la escala cognitiva y de motricidad fina, y el WHO-Motor muestra correlaciones bajas pero significativas con la escala cognitiva y de lenguaje expresivo.

3.4.4. Robustez

Antes de continuar con el análisis, examinamos la robustez de los hallazgos presentados hasta ahora al uso de las puntuaciones compuestas del Bayley-III (es decir, las puntuaciones estandarizadas externamente). Las puntuaciones compuestas del Bayley-III combinan el lenguaje receptivo y expresivo en una misma escala de lenguaje, y la motricidad fina y gruesa en una misma escala de motricidad. Por lo tanto, correlacionamos las escalas de las pruebas cortas con las tres puntuaciones compuestas del Bayley-III: cognitiva, de lenguaje y de motricidad. Los resultados que aparecen en la Tabla A2 del Apéndice III muestran que los hallazgos son por lo general similares a aquellos que se reportaron anteriormente (Tabla 6), con unas pocas diferencias en los niños más pequeños. Más concretamente, en niños menores de 19 meses, las correlaciones entre las escalas de motricidad gruesa de las pruebas cortas y la escala de motricidad del Bayley-III ya no presentan valores altos sino moderados. Es muy probable que esto se deba a que ambas escalas de motricidad, fina y gruesa, están combinadas en las puntuaciones compuestas de motricidad del Bayley-III. En este grupo etario, la correlación con la escala cognitiva del Bayley-III para la escala de motricidad fina-adaptativa del Denver-II también es más baja que la observada anteriormente (y trivial, *i. e. r* <0,20). En niños de 19 meses o más, el patrón de correlaciones es muy similar al que se obtuvo utilizando las puntuaciones estandarizadas internamente mediante métodos no paramétricos.

Los resultados también son similares, cualitativamente, a las siguientes pruebas de robustez (resultados disponibles bajo solicitud): (i) estandarizar internamente los puntajes brutos entre intervalos de 2 meses de edad, cada uno con 25 a 51 niños, siguiendo los procedimientos de estandarización paramétrica; (ii) utilizar la versión original de seis ítems del ASQ-3; (iii) dividir la muestra en grupos de 6 meses de edad, en lugar de en 12 meses de edad; (iv) controlar por prematuridad antes de estandarizar las puntuaciones; (v) calcular las correlaciones de rango de Spearman, y (vi) calcular correlaciones canónicas. Los resultados obtenidos utilizando correlaciones de Spearman o canónicas son similares y en todo caso algo mayores, lo que reforzó los hallazgos principales.

3.4.5. Correlaciones por estatus socioeconómico

Cuando se evalúa el desarrollo infantil temprano en el marco de una evaluación de impacto de los programas destinados a familias vulnerables y de escasos recursos, muchas veces se plantea si los niveles bajos de educación de la madre, que por lo general presentan estas poblaciones, afectan la calidad de los datos recogidos. Es posible que las madres que cuenten con un mejor nivel educativo estén más familiarizadas con los hitos del desarrollo, presten mayor atención a las habilidades del niño o puedan ofrecer información más precisa

sobre lo observado. Por consiguiente, la capacidad de una prueba para medir el desarrollo puede disminuir en función de la capacidad que posea el cuidador para reportar las habilidades del niño y más aún en función de la cantidad de ítems que se obtengan por reporte del cuidador principal. Es posible que mientras más ítems deba reportar el cuidador, menos confiable sea la prueba o menor sea su capacidad de medir el desarrollo.

Este estudio aporta un marco apropiado para analizar empíricamente en qué medida varía la capacidad relativa de las pruebas cortas para medir el desarrollo infantil de acuerdo con las características del cuidador principal. Consideramos la riqueza del hogar, asociada a su vez con la educación del cuidador (del padre/madre) (r =0,47), una buena *proxy* de la capacidad de este para reportar correctamente el nivel de desarrollo del niño; y repetimos el análisis de la validez concurrente para los hogares situados dentro del 25% más bajo y del 25% más alto en la distribución de riqueza del hogar de la muestra.[18] Evaluamos a través de métodos *bootstrap* si el tamaño de la correlación entre el 25% de los hogares más pobres y el 25% de los más ricos presenta una diferencia estadísticamente significativa. Para asegurarnos de que contamos con suficiente poder muestral, llevamos a cabo este análisis únicamente para la totalidad de la muestra (niños de 6 a 42 meses).

La Tabla 7, similar a la Tabla 5, presenta la validez concurrente por dominio en los niños de 6 a 42 meses dentro del 25% más bajo de la distribución de riqueza en el hogar (columnas de la izquierda) y dentro del 25% más rico (columnas de la derecha). Una comparación de los coeficientes de correlación de Pearson entre los dos tipos de hogares no revela ningún patrón específico, ni tampoco señala una diferencia en los tamaños de correlación: la diferencia en los coeficientes de correlación oscila entre 0,010 y 0,230 en valores absolutos; siendo los coeficientes mayores para los hogares más ricos en algunas ocasiones y menores en otras. Nótese que, dada la naturaleza de los resultados, limitamos este análisis a la comparación de correlaciones entre dominios del desarrollo coincidentes. Esto nos proporciona un total de 18 comparaciones. En cada una de ellas, el test de significancia estadística no permite rechazar la hipótesis nula de no diferencia, lo que indicaría que los reportes del cuidador no son sistemáticamente diferentes entre cuidadores de distinto estatus socioeconómico.

3.5. Capacitación de las pruebas y costos administrativos

La Tabla 1, que ya hemos introducido en la Sección 2.2, indica los costos de capacitación y administración de las pruebas, incluyendo el costo de la prueba por niño, y el tiempo y grado de dificultad de su administración y capacitación. Como se mencionó anteriormente, la valoración acerca del grado de dificultad de la capacitación y administración de la prueba lo reporta el capacitador y, por lo tanto, está sujeto a un cierto grado de subjetividad. Del mismo modo, la cantidad de días requeridos para la capacitación se basan en la experiencia que hemos adquirido en este y en otros estudios, y puede variar en función de la formación académica y experiencia previa tanto de los capacitadores como de los encuestadores/evaluadores. El tiempo de administración de la prueba *en este estudio* corresponde a la duración promedio registrada por el capacitador durante las administraciones que fueron supervisadas durante la recolección de datos.

En general, la tabla muestra que los costos del set de materiales y de administración por niño (costo unitario de la hoja de respuesta) son considerablemente más elevados para el Bayley-III que para el resto de las pruebas cortas. Además, el Bayley-III requiere más

[18] En lugar de utilizar educación de la madre, utilizamos riqueza del hogar debido a que su distribución presenta una mayor variabilidad, es más continua, y se distribuye de modo habitual.

tiempo para su capacitación y administración (83 minutos, en promedio) y necesita que el evaluador adquiera un mayor número de habilidades para poder administrarla. Dentro de las pruebas cortas multidimensionales, el BDI-2 es la prueba que demanda más tiempo para su administración (59 minutos, en promedio) y es también la más cara. El Denver-II y el ASQ-3 se encuentran en una posición intermedia respecto tanto del tiempo como del costo de administración. El ASQ-3 dura entre 7 y 8 minutos menos que el Denver-II, en promedio, y es un poco más caro, pero a diferencia del Denver-II incluye hojas de respuesta fotocopiables.[19] Tal y como se esperaba, las pruebas de un solo dominio del desarrollo son las más cortas y las más económicas. La administración de estas pruebas tiene una duración inferior a 10 minutos cada una, y el WHO-Motor se puede obtener sin cargo.

Por otra parte, el tiempo de administración aumenta según la longitud de la prueba y el número de ítems que se puntúen a través de administración directa o por observación de las habilidades del niño durante la evaluación. Según la valoración del capacitador, aquellas escalas o pruebas que utilizan en su mayoría reportes del cuidador principal para recolectar los datos son más fáciles de capacitar y de administrar. Las más sencillas de todas son las listas de vocabulario de los SF. Dato curioso: cuando preguntamos cuál era la prueba que más les había gustado, la respuesta de los cuidadores principales fue el WHO-Motor y el Bayley-III, que son las pruebas que requieren una menor contribución (casi ninguna) por parte del cuidador. Las escalas que administramos de estas pruebas se recogieron exclusivamente mediante evaluación directa del niño.

4. Conclusiones: Selección de la prueba y consideraciones finales

Hemos estudiado el uso de tres pruebas de tamizaje multidimensionales—el ASQ-3, el Denver-II y el BDI-2 (prueba de tamizaje de Battelle)—y dos pruebas que miden un solo dominio del desarrollo infantil—los SFI (CDI de Mac-Arthur Bates) y el WHO-Motor—en una muestra de niños de 6 a 42 meses, representativa de los hogares de clase baja y media-baja de la ciudad de Bogotá (Colombia), prestando especial atención a la confiabilidad, validez y viabilidad de las pruebas cuando estas se utilizan en un contexto similar al de una evaluación a escala. Muchas de estas pruebas se han utilizado previamente en estudios a escala en países de ingresos bajos y medios (Fernald *et al.* 2012; Macours, Schady y Vakis 2012; Fernald y Hidrobo 2011).

A lo largo de este análisis, consideramos al Bayley-III nuestro patrón de referencia. Si bien este no había sido estandarizado para Colombia, hemos demostrado que es una prueba válida para esta muestra y que fue adecuado elegirlo como patrón de referencia: hemos demostrado que presenta una buena confiabilidad de test-retest y de consistencia interna, y que está relacionado con características socioeconómicas del niño, la madre y el hogar tal y como se esperaba. De hecho, en un estudio anterior, habíamos reportado que las puntuaciones de las pruebas exhibían diferencias por cuartiles de riqueza desde el primer año de vida, que crecieron hasta los 42 meses en esta muestra (Rubio-Codina *et al.* 2015). Las escalas también mostraron niveles de validez predictiva aceptables para los dominios cognitivo, lenguaje y preparación escolar a los 5 años en un estudio contemporáneo en Colombia llevado a cabo por los mismos investigadores (comunicación personal). Además, las correlaciones de las escalas del Bayley-III entre sí también son similares a las que figuran en el manual de la prueba (Bayley 2006).

[19] La comercialización del Denver-II se suspendió recientemente. Aun así, los materiales de la prueba (manual y hojas de respuesta) se pueden descargar desde el sitio web de la editorial en su versión en inglés y en español.

La confiabilidad de las pruebas cortas analizadas es en general aceptable. A pesar de que los tamaños de la muestra son muy pequeños para calcular las confiabilidades de test-retest, todas las pruebas cortas presentan valores buenos, a excepción de la escala de motricidad fina del ASQ-3 y de las subescalas de motricidad gruesa y personal-social del Denver-II. De modo similar, la confiabilidad de la consistencia interna muestra valores buenos o aceptables, excepto en el caso del ASQ-3 y en algunas escalas del BDI-2, en particular en los niños del grupo etario intermedio. Por otra parte, la consistencia interna no siempre aumenta con la edad. Las escalas de las pruebas cortas también se correlacionan entre sí, tal y como se esperaba, y obtienen valores similares a los que figuran en los manuales, en aquellos casos en que las pruebas disponen de ellos. Esto indica que existe congruencia entre las escalas y que efectivamente hay un grado de interrelación entre los dominios del desarrollo.

Con respecto a la validez, tal y como se esperaba, todas las pruebas cortas se correlacionan con un conjunto de variables socioeconómicas del niño, la madre y el hogar, a pesar de que las correlaciones son mucho más bajas que las que se observan para el Bayley-III. El patrón de validez concurrente con el Bayley-III varía según la edad y el dominio del desarrollo, en particular para las pruebas multidimensionales, las cuales además abarcan la totalidad del rango etario bajo estudio. En términos generales, la validez concurrente aumenta con la edad para las escalas cognitiva, de lenguaje y de motricidad fina. La concurrencia de estas escalas en el Denver-II y el BDI-2 es baja pero significativa en niños menores de 19 meses, moderada en niños de 19 a 30 meses y de moderada a alta en niños mayores de 30 meses. La escala de lenguaje por lo general muestra los niveles de concurrencia más elevados a partir de los 19 meses de edad. En todo el rango etario, el ASQ-3 presenta de manera consistente una validez concurrente inferior a la del Denver-II y el BDI-2 para estas escalas. De hecho, no parece ser informativa para niños menores de 19 meses. Los resultados del ASQ-3 no varían si se utiliza la versión de seis ítems de esta prueba. A excepción del BDI-2, las escalas de motricidad gruesa se comportan de manera diferente: obtienen una validez concurrente alta en menores de 19 meses que disminuye a partir de esa edad. Entre las pruebas que miden un solo dominio del desarrollo, el WHO-Motor presenta una concurrencia alta con la escala de motricidad gruesa del Bayley-III hasta los 15 meses, y la escala de lenguaje expresivo del SFII exhibe correlaciones altas entre los 19 y 30 meses. Estos resultados se mantienen al estandarizar las puntuaciones de las pruebas utilizando diferentes métodos y también al utilizar correlaciones de rango, entre otras pruebas de robustez.

El análisis de la validez concurrente por estatus socioeconómico del hogar no denota diferencias estadísticamente significativas entre el 25% de los hogares más pobres y más ricos de la muestra. Esto parece descartar cualquier preocupación relacionada con la capacidad del cuidador para reportar el desarrollo del niño (capacidad que se considera mayor en cuidadores con un mejor nivel educativo), lo que podría haber afectado el desempeño de algunas pruebas, en particular aquellas que tienen un contenido más alto de ítems administrados por reporte del cuidador.

4.1. Selección de la prueba

La elección de la prueba óptima para utilizar en un estudio depende de la disponibilidad de tiempo, fondos y evaluadores calificados—factores que suelen ser limitados en estudios a gran escala. La elección también está condicionada a la validez de la prueba, las adaptaciones que se requieran, la edad de los niños y los objetivos del estudio. Por ejemplo, los principales indicadores de interés pueden variar dependiendo de si el objetivo es

establecer un perfil de desarrollo general para una población o evaluar una intervención, así como también según el tipo de intervención que se desea evaluar.

La totalidad de las pruebas multidimensionales cubren todo el rango etario; y la validez concurrente observada para las escalas cognitiva, de lenguaje y de motricidad fina poco se diferencia entre el BDI-2 y el Denver-II. Sin embargo, la administración del BDI-2 requiere mucho más tiempo y es más cara (tanto por el costo de los materiales como por el precio unitario de las hojas de respuesta). Además, requiere más adaptaciones y traducciones, así como una mayor cantidad de tiempo para su capacitación. El Denver-II y el ASQ-3 requieren una cantidad similar de materiales para su administración, y en ambas pruebas el tiempo de administración es menor que en el BDI-2. No obstante, dado que el ASQ-3 presenta la validez concurrente más baja en todas las escalas, parecería que el Denver-II es la prueba más apropiada—confiable, válida y viable—para usar a escala. De hecho, el bajo nivel de validez que exhibe el ASQ-3 en menores de 30 meses es preocupante, dado que esta prueba se utiliza cada vez más en estudios a gran escala (Fernald *et al.* 2012; Martinez y Naudeau 2012). A pesar de que es posible que las adaptaciones al español hayan cambiado las propiedades psicométricas del ASQ-3, los resultados de este estudio no indican nada acerca de su validez como prueba de tamizaje para niños con alto riesgo de retrasos en el desarrollo. Un dato importante para tener en cuenta es que una versión del Denver-II, adaptada para Nicaragua y administrada en el hogar, mostró sensibilidad al impacto de un programa de transferencias, lo cual respalda nuestro resultado de usar el Denver-II para la evaluación de programas a escala (Macours, Schady y Vakis 2012).

Las pruebas de un solo dominio son las más factibles de administrar, ya que son cortas, económicas y requieren menos capacitación. A pesar de que su rango etario es limitado, estas ofrecen niveles de concurrencia razonables para los dominios y edades para los que están disponibles. Por consiguiente, podrían ser consideradas en estudios a escala. El WHO-Motor muestra una validez alta para la escala de motricidad gruesa en niños menores de 16 meses. Además, está también correlacionada con las escalas cognitiva y de lenguaje expresivo, si bien las correlaciones son bajas. Esto coincide con los hallazgos del estudio bangladesí mencionado anteriormente (Hamadani *et al.* 2013). No obstante, es importante tener en cuenta que en el estudio bangladesí se realizaron evaluaciones mensuales, lo que podría ser más apropiado que una única evaluación del desarrollo motor grueso, como se realizó en este estudio.

De modo similar, las escalas de lenguaje expresivo del SFI y SFII tienen una validez al menos igual de buena que las escalas de lenguaje de las pruebas multidimensionales y exhiben correlaciones bajas con las escalas cognitiva y de motricidad fina en niños menores de 19 meses en Bogotá. En Bangladesh, los reportes de vocabulario desarrollados localmente a partir del SFII también obtuvieron una validez concurrente moderada (r=0,41, P <0,01) con el MDI del Bayley-II a los 18 meses, y predijeron el coeficiente intelectual a los 5 años (r=0,37, P <0,01) (Hamadani *et al.* 2010). Una de las ventajas de los reportes maternos sobre los niveles de vocabulario temprano adquirido es que los niños pequeños que viven en condiciones desfavorables en países de ingresos bajos y medios, quienes suelen sentirse cohibidos ante este tipo de evaluaciones, no necesitan hablar con el evaluador. Una desventaja es que para cada nuevo idioma se debe desarrollar un nuevo inventario, lo que lleva tiempo y requiere capacitación.[20] Además, es posible que se deban realizar algunas adaptaciones cuando se utiliza el mismo idioma en diferentes

[20] http://mb-cdi.stanford.edu/adaptations.html. El Consejo Consultivo del CDI de Stanford ofrece su ayuda y asesoramiento para llevar a cabo esta tarea, siempre que sea posible.

países/contextos (adaptaciones dialectales). Sin embargo, esto es factible y, de hecho, los SFs ya se encuentran disponibles en muchos idiomas.

Existe un acuerdo generalizado de que las pruebas multidimensionales son preferibles a las pruebas unidimensionales (Fernandes *et al.* 2014). Sin embargo, cuando los recursos son limitados, puede resultar adecuado (o conveniente) utilizar ciertas escalas de una prueba o pruebas de un solo dominio, dependiendo de la edad de los niños y el objetivo de la evaluación. Por ejemplo, para evaluar programas de estimulación psicosocial, que rara vez están destinados a promover el desarrollo motor grueso, se podrían utilizar las escalas de lenguaje y de motricidad fina-adaptativa del Denver-II. Sin embargo, para intervenciones nutricionales, que es más probable que tengan un impacto en el desarrollo motor en niños pequeños, el WHO-Motor podría ser de gran utilidad en niños menores de 16 meses; y más aún, dado que esta prueba posee correlaciones bajas pero significativas con las escalas cognitiva y de lenguaje expresivo. Si por el contrario, el foco de interés es el lenguaje y la muestra está constituida por niños menores de 30 meses, se podría recurrir al SFI y al SFII, descartando la escala de lenguaje receptivo.

En general, el hecho de que la validez concurrente de todas las pruebas sea de baja a moderada en los niños más pequeños, excepto para las escalas de motricidad gruesa, coincide con las dificultades reportadas sobre la medición del desarrollo infantil temprano, en particular en estudios a escala (Frongillo *et al.* 2014; Fernald *et al.* 2009; Fernandes *et al.* 2014). Como resultado, a excepción del área de desarrollo motor, todas las otras pruebas parecen tener una validez limitada en niños menores de 18 meses. No obstante, es necesario llevar a cabo un análisis de validez predictiva que complemente los hallazgos encontrados en este estudio para estar completamente seguros, dado que la validez concurrente y la validez predictiva no siempre están estrechamente relacionadas. Por ejemplo, la prueba de lenguaje en el estudio bangladesí obtuvo una validez concurrente moderada con el Bayley-II a los 18 meses y una validez predictiva similar del coeficiente intelectual en niños más grandes (Hamadani *et al.* 2010). Del mismo modo, sería conveniente llevar a cabo más investigaciones acerca del desarrollo de nuevas pruebas y posibles modificaciones a pruebas existentes para niños menores de 24 meses.

4.2. Consideraciones finales

La medición del DIT en niños muy pequeños a escala es un desafío. No obstante, las pruebas de tamizaje multidimensionales y las pruebas que miden un solo dominio son alternativas viables y confiables. La validez concurrente varía según el dominio y el rango etario. Las escalas que obtuvieron los valores más altos de validez concurrente son las escalas de motricidad gruesa en menores de 19 meses y las escalas de lenguaje en niños mayores que esa edad. El Denver-II es la prueba multidimensional más válida y factible de administrar, y el ASQ-3 presenta por lo general un desempeño pobre en niños menores de 31 meses. Es necesario investigar la validez predictiva y el grado de sensibilidad de estas pruebas ante distintas intervenciones para respaldar aún más los resultados de este estudio. Esto sería de gran utilidad para la selección de instrumentos de medición y para el diseño de estudios futuros que requieran medir el desarrollo infantil a gran escala. Actualmente, estamos planificando un estudio de seguimiento con el objetivo de estudiar la validez predictiva relativa de las pruebas cortas y el Bayley-III respecto del coeficiente intelectual, el desarrollo del lenguaje, la función ejecutiva y el rendimiento académico, en niños de 6 a 9 años. Los resultados que surjan de este estudio de seguimiento complementarán los hallazgos reportados en este artículo.

Referencias

Attanasio, O. P., C. Fernandez, E. O. A. Fitzsimons, S. M. Grantham-McGregor, C. Meghir, y M. Rubio-Codina. 2014. “Using the Infrastructure of a Conditional Cash Transfer Program to Deliver a Scalable Integrated Early Child Development Program in Colombia: Cluster Randomized Controlled Trial.” *BMJ* 349 (sep29 5): g5785–g5785.

Attanasio, O. P. 2015. “The Determinants of Human Capital Formation during the Early Years of Life: Theory, Measurement, and Policies.” *Journal of the European Economic Association* 13 (6): 949–97.

Bayley, N. 1969. *Bayley Scales of Infant Development*. New York: Psychological Corp.

———. 2006. *Bayley Scales of Infant and Toddler Development–Third Edition: Technical Manual.* San Antonio, TX: Harcourt Assessment.

Campbell, F., G. Conti, J. J. Heckman, S. H. Moon, R. Pinto, E. Pungello, y Y. Pan. 2014. “Early Childhood Investments Substantially Boost Adult Health.” *Science* 343 (6178): 1478–85.

Efron, B. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans*. *CBMS-NSF Regional Conference Series in Applied Mathematics*. Vol. 38. Philadelphia,PA: SIAM.

Engle, P. L., M. M. Black, J. R. Behrman, M. Cabral de Mello, P. J. Gertler, L. Kapiriri, R. Martorell, y M. E. Young. 2007. “Strategies to Avoid the Loss of Developmental Potential in More than 200 Million Children in the Developing World.” *Lancet* 369 (9557): 229–42.

Evans, J. D. 1996. *Straightforward Statistics for the Behavioral Sciences*. Edited by Brooks/Cole Publishing. Pacific Grove, CA.

Fenson, L., P. S. Dale, J. S. Reznick, D. Thal, E. Bates, J. P. Hartung, S. J. Pethick, y J. S. Reilly. 2002. *The MacArthur Communicative Development Inventories: Users Guide and Technical Manual*. Baltimore, MD: Paul Brookes Publishing Co.

Fernald, L. C., P. Kariger, M. Hidrobo, y P. J. Gertler. 2012. “Socioeconomic Gradients in Child Development in Very Young Children: Evidence from India, Indonesia, Peru, and Senegal.” *Proceedings of the National Academy of Sciences* 109 (Supplement_2): 17273–80.

Fernald, L. C., P. Kariger, P. L. Engle, y A Raikes. 2009. “Examining Child Development in Low-Income Countries: A Toolkit for the Assessment of Children in the First Five Years of Life.” Washington, D.C.

Fernald, L. C. y M. Hidrobo. 2011. “Effect of Ecuador’s Cash Transfer Program (Bono de Desarrollo Humano) on Child Development in Infants and Toddlers: A Randomized Effectiveness Trial.” *Social Science and Medicine* 72 (9): 1437–46.

Fernald, L C., A. Weber, E. Galasso, y L. Ratsifandrihamanana. 2011. “Socioeconomic Gradients and Child Development in a Very Low Income Population: Evidence from Madagascar.” *Developmental Science* 14 (4): 832–47.

Fernandes, M., A. Stein, C. R. Newton, L. Cheikh-Ismail, M. Kihara, K. Wulff, E. de León Quintana, *et al*. 2014. “The INTERGROWTH-21st Project Neurodevelopment Package: A Novel Method for the Multi-Dimensional Assessment of Neurodevelopment in Pre-School Age Children.” *PloS One* 9 (11): e113360.

Frankenburg, W. K., J. Dodds, P. Archer, B. Bresnick, P. Maschka, N. Edelmann, y H.

Shapiro. 1990. *The DENVER II Technical Manual*. Denver, CO: Denver Developmental Materials.

Frankenburg, W. K., J. Dodds, P. Archer, H. Shapiro, y B. Bresnick. 1992. “A Major Revision and Restandardization of the Denver Developmental Screening Test.” *Pediatrics* 89: 91–97.

Frongillo, E. A., F. Tofail, J. D. Hamadani, A. M. Warren, y S. F. Mehrin. 2014. “Measures and Indicators for Assessing Impact of Interventions Integrating Nutrition, Health, and Early Childhood Development.” *Annals of the New York Academy of Sciences* 1308 (1): 68–88.

Frongillo, E. A., S. M. Sywulka, y P. Kariger. 2003. “UNICEF Psychosocial Care Indicators Project.”

Gertler, P., J. J Heckman, R. Pinto, A. Zanolini, C. Vermeersch, S. Walker, S. M. Chang, y S. M. Grantham-McGregor. 2014. “Labor Market Returns to an Early Childhood Stimulation Intervention in Jamaica.” *Science* 344 (6187): 998–1001.

Grantham-McGregor, S., Y. Bun Cheung, S. Cueto, P. Glewwe, L. Richter, y Barbara Strupp. 2007. “Developmental Potential in the First 5 Years for Children in Developing Countries.” *Lancet* 369 (9555): 60–70.

Greenspan, S. I. 2004. *Greenspan Social-Emotional Growth Chart: A Screening Questionnaire for Infants and Young Children*. San Antonio, TX: Harcourt Assessment, Inc.

Hamadani, J. D., F. Tofail, S. N. Huda, D. S. Alam, D. a. Ridout, O. Attanasio, y S. M. Grantham-McGregor. 2014. “Cognitive Deficit and Poverty in the First 5 Years of Childhood in Bangladesh.” *Pediatrics* 134 (4): e1001–8.

Hamadani, J. D., S. N. Huda, F. Khatun, y S. M Grantham-McGregor. 2006. “Psychosocial Stimulation Improves the Development of Undernourished Children in Rural Bangladesh.” *The Journal of Nutrition* 136 (10): 2645–52.

Hamadani, J. D., H. Baker-Henningham, F. Tofail, F. Mehrin, S. N. Huda, y S. M. Grantham-McGregor. 2010. “Validity and Reliability of Mothers’ Reports of Language Development in 1-Year-Old Children in a Large-Scale Survey in Bangladesh.” *Food and Nutrition Bulletin* 31 (2 SUPPL.): 198–206.

Hamadani, J. D., F. Tofail, T. Cole, y S. Grantham-McGregor. 2013. “The Relation between Age of Attainment of Motor Milestones and Future Cognitive and Motor Development in Bangladeshi Children.” *Maternal and Child Nutrition* 9 (SUPPL. 1): 89–104.

Harrison, P. L.,y T Oakland. 2003. *Adaptive Behavior Assessment System-Second Edition.* San Antonio, TX: The Psychological Corporation.

Heckman, J. J. 2007. “The Economics, Technology, and Neuroscience of Human Capability Formation.” *Proceedings of the National Academy of Sciences* 104 (33): 13250–55.

Jackson-Maldonado, D., V. A. Marchman, y L. C. Fernald. 2012. “Short-Form Versions of the Spanish MacArthur–Bates Communicative Development Inventories.” *Applied Psycholinguistics*, 1–32.

Jackson-Maldonado, D., D. Thal, V. Marchman, T. Newton, L. Fenson, y B. Conboy. 2003. *Mac Arthur Inventarios Del Desarrollo de Habilidades Comunicativas. User’s Guide and Technical Manual*. Baltimore, MD: Paul H. Brookes Publishing Co.

Luby, J. L. 2015. “Poverty ’ S Most Insidious Damage: The Developing Brain.”*JAMA Pediatrics*, 1–2.

Macours, K., N. Schady, y R. Vakis. 2012. “Cash Transfers, Behavioral Changes, and Cognitive Development in Early Childhood: Evidence from a Randomized Experiment.” *American Economic Journal: Applied Economics* 4 (2): 247–73.

Martinez, S. y S. Naudeau. 2012. “The Promise of Preschool in Africa : A Randomized Impact Evaluation of Early Childhood Development in Rural Mozambique.”

Nahar, B., J. D. Hamadani, T. Ahmed, F. Tofail, A. Rahman, S. N. Huda, y S. M. Grantham-McGregor. 2009. “Effects of Psychosocial Stimulation on Growth and Development of Severely Malnourished Children in a Nutrition Unit in Bangladesh.” *European Journal of Clinical Nutrition* 63 (6). Nature Publishing Group: 725–31.

Newborg, J. 2005. *Battelle Developmental Inventory--Second Edition*. Itasca, IL: Riverside Publishing.

Rubio-Codina, M., O. Attanasio, y S. Grantham-McGregor. 2015. “Mediating Pathways in the Socioeconomic Gradient of Child Development: Evidence from Children 6-42 Months in Bogota.” *International Journal of Behavioral Development, June 2016.* doi: 10.1177/0165025415626515.

Rubio-Codina, M., O. Attanasio, C. Meghir, N. Varela, y S. Grantham-McGregor. 2015. “The Socio-Economic Gradient of Child Development Children 6-42 Months in Bogota.” *The Journal of Human Resources* 50 (2): 464–83.

Schady, N., J. Behrman, M. C. Araujo, R. Azuero, R. Bernal, D. Bravo, F. Lopez-Boo, *et al.* 2015. “Wealth Gradients in Early Childhood Cognitive Development in Five Latin American Countries.” *The Journal of Human Resources* 50 (2): 446–63.

Squires, J., D. Bricker, E. Twombly, R. Nickel, J. Clifford, K. Murphy, R. Hoselton, L. Potter, L. Mounts, y J. Farrell. 2009. *Ages & Stages English Questionnaires, Third Edition (ASQ-3): A Parent-Completed, Child-Monitoring System*. Baltimore, MD: Paul H. Brookes Publishing Co.

UN General Assembly. 2015. “Resolution Adopted by the General Assembly on 25 September 2015.”

Verdisco, A., S. Cueto, J. Thompson, P. Engle, O. Neuschmidt, S. Meyer, E. González, B. Oré, K. Hepworth, y A. Miranda. 2009. “Urgency and Possibility Results of PRIDI A First Initiative to Create Regionally Comparative Data on Child Development in Four Latin American Countries Technical Annex.”

Walker, S. P, S. M. Chang, M. Vera-Hernández, y S. Grantham-McGregor. 2011. “Early Childhood Stimulation Benefits Adult Competence and Reduces Violent Behavior.” *Pediatrics* 127 (5): 849–57.

Walker, S. P., T. D. Wachs, S. Grantham-Mcgregor, M. M. Black, C. A. Nelson, S. L. Huffman, H. Baker-Henningham, *et al.* 2011. “Inequality in Early Childhood: Risk and Protective Factors for Early Child Development.” *The Lancet* 378 (9799): 1325–38.

WHO Multicentre Growth Reference Study Group. 2006. “WHO Growth Standards: Length/height-for-Age, Weight-for-Age, Weight-for-Length, Weight-for-Height and Body Mass Index-for-Age: Methods and Development.” Geneva.

WHO Multicentre Growth Reference Study Group. 2006. “WHO Motor Development Study: Windows of Achievement for Six Gross Motor Development Milestones.” *Acta Paediatrica. Supplementum* 450: 86–95.

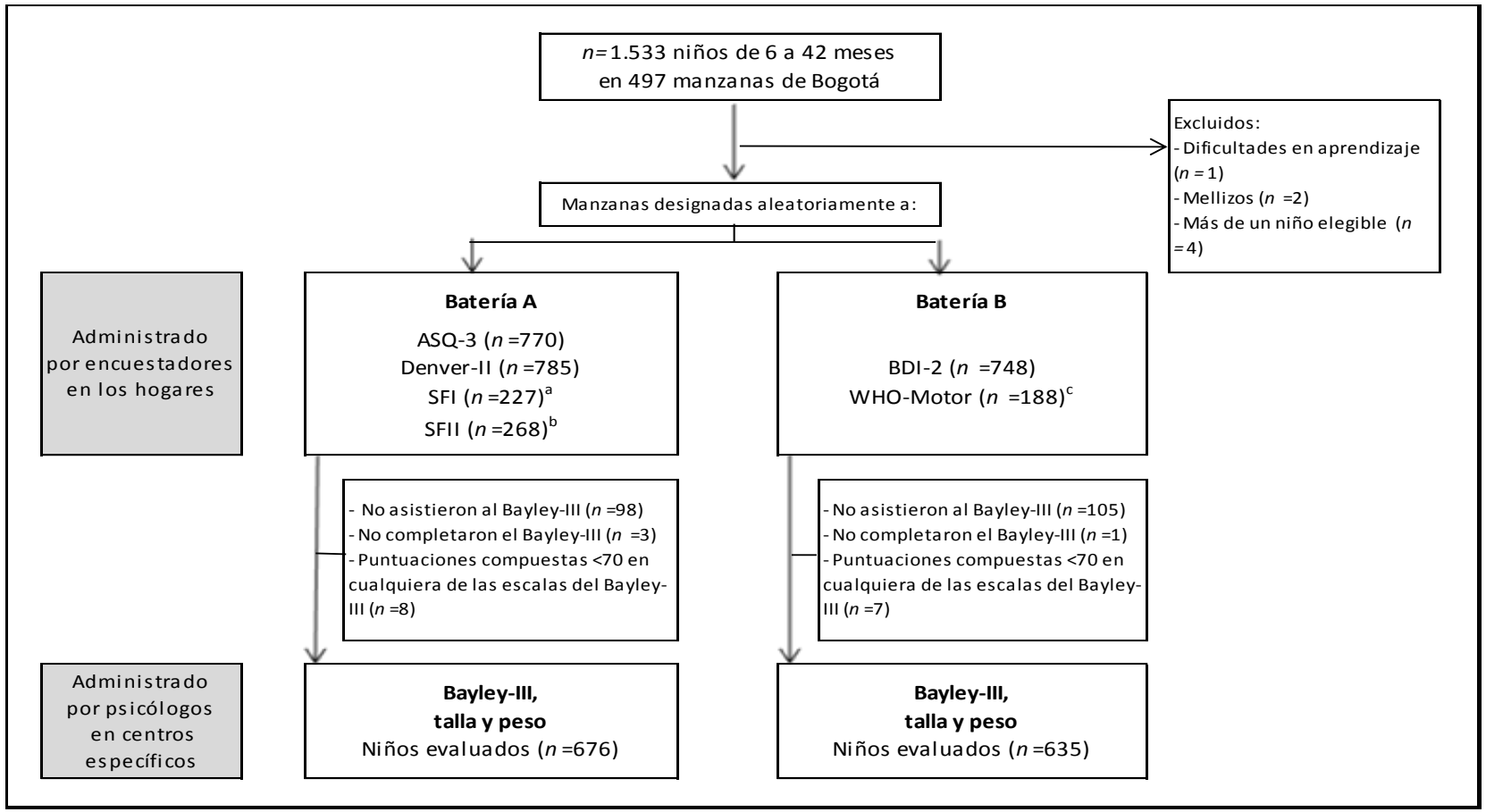
World Health Organization. 1983. “Measuring Change in Nutritional Status. Guidelines for Assessing the Nutritional Impact of Supplementatry Feeding Programmes for Vulnerable Groups.” Geneva.

Wijnhoven, T. M. A., M. de Onis, A. W. Onyango, T. Wang, G. A. Bjoerneboe, N. Bhandari, A. Lartey, y B. Al Rashidi. 2004. “Asessment of Gross Motor Development in the WHO Multicentre Growth Reference Study.” *Food and Nutrition Bulletin* 25 (1 SUPPL. 1).

Wolf, S., P. Halpin, H. Yoshikawa, L. Pisani, A. J. Dowd, e I. Borisova. 2016 "Assessing the construct validity of Save the Children's International Development and Early Learning Assessment (IDELA)." mimeo.

Tablas y gráficos

Gráfico 1: Diagrama de flujo de participantes en el estudio



[a] Niños de 8 a 18 meses.

[b] Niños de 19 a 30 meses.

[b] Niños de 6 a 15 meses.

Gráfico 2: Validez concurrente entre el Bayley-III y las escalas coincidentes en las pruebas cortas por dominio y grupo etario

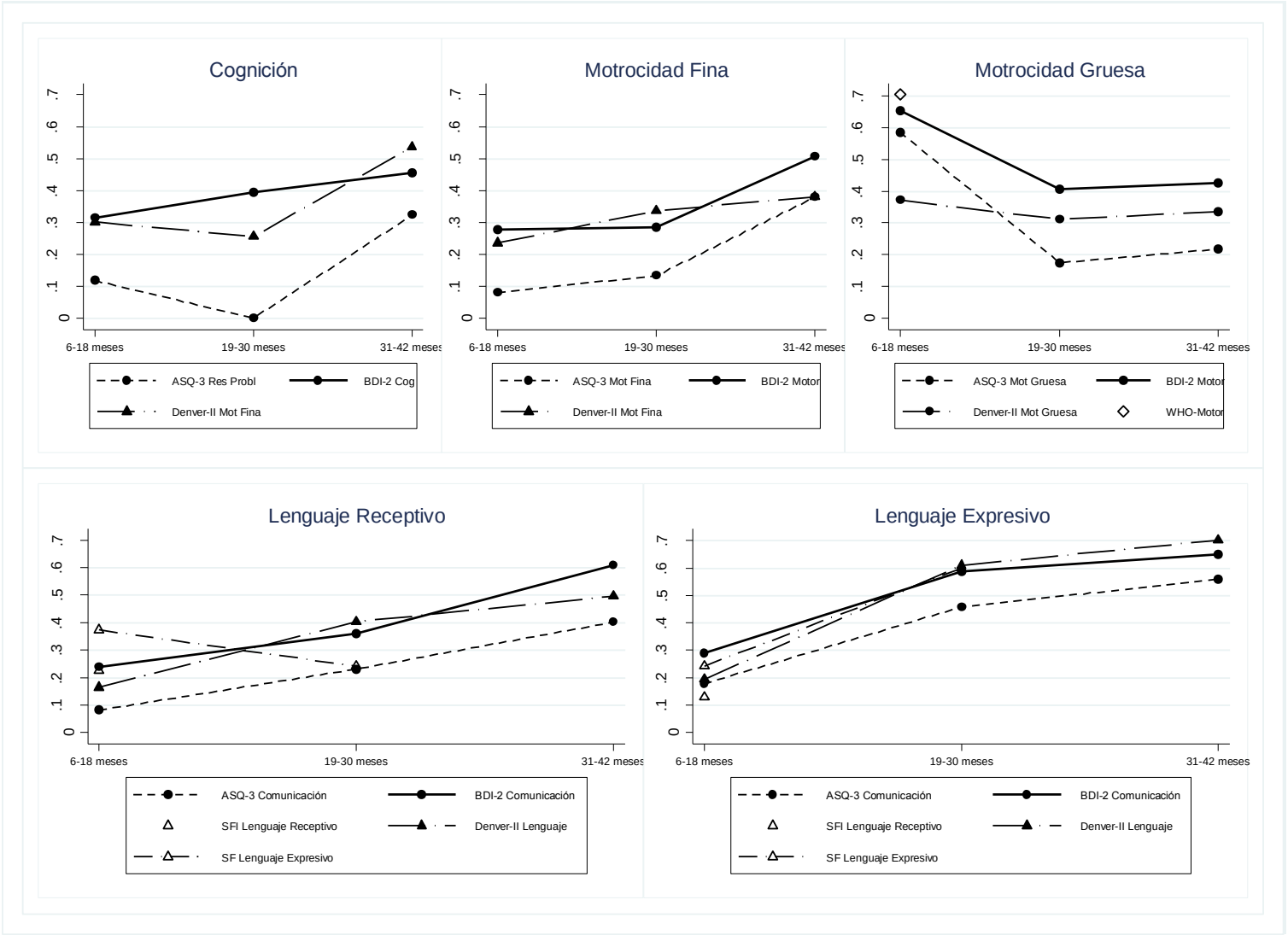


Tabla 1: Características del Bayley-III y las pruebas cortas

Prueba	Escalas incluidas en el estudio y cantidad de ítems[a]	Rango etario de la prueba (meses)	Rango etario en la muestra (meses)	Costo (USD)[b]	Minutos para la adm. [c]	Minutos para la adm. en el estudio[d]	Días de capacitación[e]	Dificultad en la capacitación[f]	Dificultad en la administración[f]
Bayley-III	Cognitiva (91) [21] Lenguaje receptivo (48) [18] Lenguaje expresivo (49) [16] Motricidad fina (66) [18] Motricidad gruesa (72) [16]	0-42	6-42	$1.050 kit + $9,34 p.nñ	30-95	(*n* =36) 83,2 (18,8)	15 + práctica	alta	alta
ASQ-3	Resolución de problemas (6) [6.5] Comunicación (6) [6.5] Motricidad fina (6) [6.4] Motricidad gruesa (6) [6.6] Personal-social (6) [6.4]	1-66	6-42	$275 kit $295 materiales	10-15	(*n* =32) 19,7 (8,2)	6 + práctica	media	media
Denver-II	Motricidad fina-adaptativa (29) [9] Lenguaje (39) [10] Motricidad gruesa (32) [9] Personal-social (25) [9]	0-71	6-42	$200 kit + $0,45 p.nñ	15-20	(*n* =32) 27 (10,5)	7 + práctica	alta	media/ alta
SFI (MacArthur)	Lenguaje receptivo (104) [104] Lenguaje expresivo (104) [104]	8-18	8-18	$90 kit + $1 p.nñ	10	(*n* =8) 8,6 (1,9)	0,5 + mínimo de práctica	baja	baja
SFII (MacArthur)	Lenguaje expresivo (100) [100]	19-30	19-30		10	(*n* =10) 8,2 (3,3)			
BDI-2 (Battelle)	Cognitiva (20) [9] Comunicación (20) [9] Motricidad (20) [9] Personal-social (20) [10] Habilidades adaptativas (20) [9]	0-83	6-42	$405,70 kit + $3,08 p.nñ	10-30	(*n* =30) 59 (13,0)	8 + práctica	alta	alta
WHO-Motor	Motricidad gruesa (6) [6]	4-24	6-15	Gratis	10	(*n* =9) 6 (2,7)	1 + práctica	media	media

[a] Cantidad total de ítems en la escala entre paréntesis y promedio de ítems evaluados por participante en el estudio entre corchetes.

[b] La información respecto de los costos se consultó por última vez en marzo de 2016 desde los sitios web de las editoriales (para obtener más detalles véase Apéndice I). 'P.nñ' corresponde a costo de administración 'por niño'. Los kits incluyen hojas de respuesta en paquetes de 100 para el Denver-II, de 30 para el BDI-2 y de 25 para el resto de las pruebas cortas y el Bayley-III. El WHO-Motor no estaba disponible en español y solo estaban disponibles algunas partes del BDI-2. La versión en español del Bayley-III se encuentra disponible desde mediados de 2015. El resto de las pruebas y manuales estaban disponibles en español.

[c] Tiempo de administración reportado en el sitio web de la editorial.

[d] Los datos son el promedio (desviación estándar) en minutos, según lo registrado por el capacitador durante las actividades de supervisión en campo.

[e] La cantidad de días son aproximaciones basadas en nuestra experiencia y están sujetas a cambios dependiendo de la formación académica y experiencia previa de los capacitadores y encuestadores/evaluadores, entre otros factores.

[f] Según lo reportado por el capacitador.

Tabla 2: Características de los niños de la muestra y sus familias por batería

	Batería A (n_A=676)	Batería B (n_B =635)	Diferencia en el valor *p* entre las baterías
I. Características del niño			
Edad del niño, %			
6-18 meses	33,7	33,9	0,960
19-30 meses	33,6	35,7	0,410
31-42 meses	32,7	30,4	0,371
Niñas, %	47,6	51,0	0,220
Prematuro (edad gestacional <37 semanas), %	15,2	15,1	0,952
Peso al nacer[a], gr, promedio (DE)	3.066 (514)	3.015 (510)	0,087
Baja talla por edad (puntaje z de talla por edad <-2DE)	16,9	17,7	0,710
II. Características de los padres			
Edad de la madre[a], y, promedio (DE)	27,2 (6,9)	26,6 (6,4)	0,106
Educación de la madre[a], y, promedio (DE)	10,3 (3,4)	10,4 (3,3)	0,541
La madre superó el nivel secundario	31,0	30,1	0,729
La madre trabaja (trab. remunerado o no remunerado)	49,7	53,9	0,134
La madre dio a luz antes de los 18 años	13,3	13,5	0,903
Educación del padre[a], y, promedio (DE)	8,1 (4,0)	8,7 (4,0)	0,026
El padre superó el nivel secundario	29,1	27,6	0,629
Padre ausente (fallecido o ausente en el hogar)	32,0	33,5	0,540
III. Características del hogar			
Sector socioeconómico (estrato), %			
1 Más pobre	30,3	29,8	0,825
2	32,7	37,0	0,101
3	36,1	32,6	0,183
4 Más rico	0,9	0,6	0,592
Índice de riqueza del hogar	-0,089 (1)	0,05 (1)	0,013
Variedad en los materiales de juego	4,8 (2,3)	4,8 (2,4)	0,682
Variedad en las actividades de juego	3,7 (1,9)	3,6 (1,8)	0,235
IV. Puntuaciones compuestas del Bayley-III, promedio (DE)			
Cognitiva	58,7 (14,5)	58,8 (13,5)	0,876
Lenguaje receptivo	25,3 (9,3)	25,6 (8,9)	0,528
Lenguaje expresivo	25,0 (10,3)	25,5 (10,2)	0,352
Motricidad fina	39,5 (10,1)	39,1 (10,0)	0,442
Motricidad gruesa	52,2 (11,7)	52,6 (11,4)	0,461

[a]Datos incompletos respecto de las siguientes variables: peso al nacer (n_A=638, n_B=552), edad de la madre y situación laboral de la madre (n_A=668, n_B=618), educación de la madre (n_A=674, n_B=633), educación del padre (n_A=639, n_B=576). DE significa desviación estándar. El índice de riqueza del hogar y la variedad en los materiales y actividades de juego se calculan de acuerdo con los procedimientos descritos en el texto.

Tabla 3: Confiabilidad de test-retest y de consistencia interna (para la totalidad de la muestra y por grupo etario) del Bayley-III y las pruebas cortas

	Test-retest CCI	Alfa de Cronbach 6-42 meses	Alfa de Cronbach 6-18 meses	Alfa de Cronbach 19-30 meses	Alfa de Cronbach 31-42 meses
Puntajes brutos del Bayley-III	(*n* = 20)	(*n* = 1.311)	(*n* = 443)	(*n* = 454)	(*n* = 414)
Cognitiva	0,96	0,97	0,94	0,90	0,82
Lenguaje receptivo	0,96	0,96	0,85	0,90	0,79
Lenguaje expresivo	0,98	0,96	0,88	0,90	0,91
Motricidad fina	0,98	0,96	0,92	0,85	0,85
Motricidad gruesa	0,98	0,97	0,96	0,85	0,78
Puntuaciones compuestas del Bayley-III					
Cognitiva	0,96	0,97	0,94	0,90	0,82
Lenguaje	0,97	0,98	0,93	0,94	0,92
Motricidad	0,98	0,98	0,97	0,91	0,88
ASQ-3 (9 ítems)	(*n* = 12)	(*n* = 664)	(*n* = 221)	(*n* = 224)	(*n* = 219)
Resolución de problemas	0,80	0,60	0,54	0,62	0,66
Comunicación	0,92	0,68	0,58	0,71	0,78
Motricidad fina	0,37	0,57	0,63	0,44	0,65
Motricidad gruesa	0,90	0,70	0,76	0,66	0,68
Personal-social	0,73	0,55	0,57	0,44	0,65
ASQ-3 (6 ítems)	(*n* = 12)	(*n* = 664)	(*n* = 221)	(*n* = 224)	(*n* = 219)
Resolución de problemas	0,80	0,42	0,45	0,32	0,51
Comunicación	0,92	0,52	0,47	0,55	0,55
Motricidad fina	0,37	0,44	0,49	0,37	0,45
Motricidad gruesa	0,90	0,55	0,72	0,47	0,39
Personal-social	0,73	0,38	0,41	0,33	0,40
Denver-II	(*n* = 12)	(*n* = 658)	(*n* = 225)	(*n* = 221)	(*n* = 212)
Lenguaje	0,93	0,93	0,85	0,85	0,90
Motricidad fina-adaptativa	0,83	0,91	0,86	0,81	0,78
Motricidad gruesa	0,53	0,90	0,90	0,78	0,74
Personal-social	0,49	0,91	0,90	0,76	0,76
SFI (MacArthur)	(*n* = 12)	(*n* = 192)	(*n* = 192)		
Lenguaje receptivo	0,99	0,97	0,97		
Lenguaje expresivo		0,92	0,92		
SFII (MacArthur)		(*n* = 226)		(*n* = 226)	
Lenguaje expresivo	NA	0,98		0,98	
BDI-2 (Battelle)	(*n* = 11)	(*n* = 635)	(*n* = 215)	(*n* = 227)	(*n* = 193)
Cognitiva	0,92	0,79	0,62	0,40	0,72
Comunicación	0,94	0,89	0,76	0,67	0,78
Motricidad	0,98	0,88	0,83	0,63	0,54
Personal-social	0,71	0,84	0,73	0,65	0,76
Habilidades adaptativas	0,90	0,84	0,71	0,61	0,62
WHO-Motor	(*n* = 11)	(*n* = 152)	(*n* = 152)		
Motricidad gruesa	0,80	0,86	0,86		

Tabla 4: Correlaciones del Bayley-III y las pruebas cortas con variables socioeconómicas

	Educación de la madre	Índice de riqueza	Actividades de juego	Materiales de juego	Prematuridad	Baja talla por edad
Bayley-III (*n* =1.311)						
Cognitiva	0,210***	0,235***	0,189***	0,271***	-0,096***	-0,051
Lenguaje receptivo	0,216***	0,191***	0,214***	0,248***	-0,056*	-0,080**
Lenguaje expresivo	0,206***	0,224***	0,209***	0,243***	-0,063*	-0,069*
Motricidad fina	0,124***	0,145***	0,119***	0,179***	-0,038	-0,082**
Motricidad gruesa	0,079**	0,034	0,023	0,056*	-0,092***	-0,051
ASQ-3 (*n* =664)						
Resolución de problemas	0,127**	0,071	0,176***	0,177***	-0,008	0,002
Comunicación	0,142***	0,136***	0,222***	0,156***	-0,012	-0,009
Motricidad fina	0,063	0,067	0,167***	0,133***	0,012	-0,011
Motricidad gruesa	-0,025	0,046	0,069	0,019	0,005	-0,012
Personal-social	0,019	0,034	0,152***	0,088*	0,007	0,062
Denver-II (*n* =658)						
Lenguaje	0,170***	0,165***	0,184***	0,173***	-0,012	-0,049
Motricidad fina-adaptativa	0,102**	0,121**	0,097*	0,109**	-0,063	-0,027
Motricidad gruesa	0,022	0,020	-0,021	-0,011	0,018	-0,068
Personal-social	-0,034	-0,019	0,064	0,010	0,022	0,006
SFI (MacArthur) (*n* =192)[a]						
Lenguaje receptivo	0,147*	0,127	0,267***	0,251***	-0,049	-0,012
Lenguaje expresivo	0,040	-0,060	-0,007	-0,005	-0,092	0,024
SFII (MacArthur) (*n* =226)[b]						
Lenguaje expresivo	0,136*	0,094	0,229***	0,200**	-0,058	0,021
BDI-2 (Battelle) (*n* =635)						
Cognitiva	0,202***	0,173***	0,164***	0,181***	0,004	-0,056
Comunicación	0,210***	0,176***	0,224***	0,245***	-0,080	-0,152***
Motricidad	0,139***	0,163***	0,135***	0,179***	-0,016	-0,102*
Personal-social	0,144***	0,136***	0,240***	0,231***	-0,012	-0,057
Habilidades adaptativas	0,074	0,094*	0,276***	0,193***	-0,029	-0,123**
WHO-Motor (*n* =152)[c]						
Motricidad gruesa	-0,036	0,008	0,082	0,018	-0,103	0,116

* p<0,05, ** p<0,01, *** p<0,001. [a]Niños de 8 a 18 meses. [b]Niños de 19 a 30 meses. [c]Niños de 6 a 15 meses.

Tabla 5: Correlaciones entre las escalas del Bayley-III y entre las escalas del Bayley-III y las pruebas cortas, niños de 6 a 42 meses

	Bayley-III, 6-42 meses				
	Cognitiva	**Lenguaje receptivo**	**Lenguaje expresivo**	**Motricidad fina**	**Motricidad gruesa**
Bayley-III			*n* = 1.311		
Cognitiva	1				
Lenguaje receptivo	0,544***	1			
Lenguaje expresivo	0,483***	0,563***	1		
Motricidad fina	0,529***	0,461***	0,413***	1	
Motricidad gruesa	0,369***	0,356***	0,306***	0,380***	1
ASQ-3			*n* = 664		
Resolución de problemas	**0,146*** **	0,156***	0,221***	0,151***	0,048
Comunicación	0,199***	**0,236*** **	**0,395*** **	0,164***	0,126**
Motricidad fina	0,172***	0,157***	0,192***	**0,200*** **	0,163***
Motricidad gruesa	0,066	0,073	0,043	0,067	**0,325*** **
Personal-social	0,100*	0,134***	0,172***	0,124**	0,098*
Denver-II			*n* = 658		
Lenguaje	0,329***	**0,401*** **	**0,506*** **	0,246***	0,193***
Motricidad fina-adaptativa	**0,386*** **	0,329***	0,308***	**0,354*** **	0,210***
Motricidad gruesa	0,216***	0,234***	0,183***	0,204***	**0,499*** **
Personal-social	0,244***	0,215***	0,226***	0,195***	0,184***
SFI & SFII (MacArthur)			*n* = 418		
Lenguaje receptivo[a]	0,187**	**0,224** **	**0,130**	0,088	0,168*
Lenguaje expresivo[b]	0,206***	**0,299*** **	**0,441*** **	0,131**	0,152**
BDI-2 (Battelle)			*n* = 635		
Cognitiva	**0,363*** **	0,319***	0,337***	0,308***	0,210***
Comunicación	0,343***	**0,349*** **	**0,495*** **	0,237***	0,263***
Motricidad	0,269***	0,220***	0,252***	**0,316*** **	**0,339*** **
Personal-social	0,124**	0,161***	0,209***	0,040	0,058
Habilidades adaptativas	0,153***	0,225***	0,233***	0,190***	0,208***
WHO-Motor			*n* = 152[c]		
Motricidad gruesa	0,224**	0,126	0,282***	0,061	**0,703*** **

* p<0,05, ** p<0,01, *** p<0,001. Correlaciones de Pearson en puntuaciones estandarizadas internamente; los errores estándar (EE) se calcularon utilizando métodos *bootstrap* , estratificando por grupo etario y sector socioeconómico (*n* =1.000 repeticiones). Las correlaciones de las escalas que miden el mismo dominio del desarrollo están resaltadas en negrita. [a]Niños de 8 a 18 meses. [b]Niños de 8 a 30 meses. [c]Niños de 6 a 15 meses.

Tabla 6: Correlaciones entre las escalas del Bayley-III y entre las escalas del Bayley-III y las pruebas cortas, por grupo etario

	Bayley-III, 6-18 meses					Bayley-III, 19-30 meses					Bayley-III, 31-42 meses				
	Cognitiva	Lenguaje receptivo	Lenguaje expresivo	Motricidad fina	Motricidad gruesa	Cognitiva	Lenguaje receptivo	Lenguaje expresivo	Motricidad fina	Motricidad gruesa	Cognitiva	Lenguaje receptivo	Lenguaje expresivo	Motricidad fina	Motricidad gruesa
Bayley-III			*n* = 443					*n* = 454					*n* = 414		
Cognitiva	1					1					1				
Lenguaje receptivo	0,437***	1				0,604***	1				0,590***	1			
Lenguaje expresivo	0,356***	0,502***	1			0,494***	0,554***	1			0,603***	0,639***	1		
Motricidad fina	0,533***	0,490***	0,408***	1		0,525***	0,435***	0,377***	1		0,528***	0,458***	0,457***	1	
Motricidad gruesa	0,333***	0,329***	0,232***	0,354***	1	0,392***	0,407***	0,350***	0,421***	1	0,381***	0,329***	0,334***	0,363***	1
ASQ-3			*n* = 221					*n* = 224					*n* = 219		
Resolución de problemas	**0,119**	0,010	0,071	0,062	0,026	**0,001**	0,075	0,133*	0,091	0,005	**0,323*** **	0,374***	0,454***	0,292***	0,111
Comunicación	0,104	**0,082**	**0,178** **	0,142*	0,164*	0,141*	**0,231*** **	**0,458*** **	0,069	0,067	0,361***	**0,402*** **	**0,560*** **	0,286***	0,147*
Motricidad fina	0,084	0,084	0,077	**0,082**	0,151*	0,135*	0,131*	0,171**	**0,134***	0,176**	0,297***	0,256***	0,331***	**0,380*** **	0,164*
Motricidad gruesa	0,147*	0,148*	0,047	0,105	**0,585*** ** [b]	-0,053	0,013	0,025	0,010	**0,175** **	0,110	0,060	0,059	0,090	**0,218*** **
Personal-social	0,208**	0,184**	0,166*	0,208**	0,070	0,033	0,082	0,121	0,065	0,109	0,063	0,140*	0,240***	0,102	0,119
Denver-II			*n* = 225					*n* = 221					*n* = 212		
Lenguaje	0,206**	**0,238*** ** [a]	**0,290*** **	0,187*	0,125	0,224***	**0,361*** ** [a,d]	**0,587*** ** [a]	0,111	0,175**	0,560*** [e]	**0,608*** ** [a]	**0,650*** **	0,443***	0,284***
Motricidad fina-adaptativa	**0,315*** ** [a]	0,279***	0,168*	**0,277*** **	0,153*	**0,395*** ** [a]	0,339***	0,345***	**0,286*** ** [a]	0,257***	**0,455*** **	0,377***	0,426***	**0,507*** **	[a] 0,229***
Motricidad gruesa	0,264***	0,270***	0,085	0,180**	**0,654*** ** [b]	0,133*	0,171**	0,207***	0,198**	**0,406*** ** [a]	0,256***	0,263***	0,270***	0,239***	**0,426*** ** [a]
Personal-social	0,366***	0,279***	0,296***	0,240***	0,244***	0,099	0,174**	0,185**	0,182**	0,197**	0,274***	0,194**	0,200**	0,166*	0,111
SFI & SFII (MacArthur)			*n* = 192 [f]					*n* = 226							
Lenguaje receptivo	0,187**	**0,224*** **	**0,130**	0,088	0,168*										
Lenguaje expresivo	0,258***	**0,373*** ** [a, b, c]	**0,242*** **	0,204***	0,176*	0,168*	**0,241*** **	**0,600*** ** [a]	0,077	0,134*					
BDI-2 (Battelle)			*n* = 215					*n* = 227					*n* = 193		
Cognitiva	**0,302*** ** [a]	0,223***	0,209**	0,274***	0,244***	**0,256*** ** [a]	0,297***	0,327***	0,253***	0.146	**0,536*** ** [a]	0,444***	0,484***	0,404***	0,243***
Comunicación	0,205**	**0,164***	**0,194** **	0,195**	0,229***	0,350***	**0,403*** ** [a]	**0,610*** ** [a]	0,196**	0,245***	0,488***	**0,496*** **	**0,702*** **	[a] 0,329***	0,325***
Motricidad	0,147*	0,161*	0,109	**0,236*** **	**0,371*** **	0,288***	0,231***	0,268***	**0,337*** ** [a]	**0,311*** **	0,379***	0,273**	0,386***	**0,380*** **	**0,335*** **
Personal-social	0,025	0,072	0,057	0,001	0,077	0.145	0,193**	0,286***	0,019	-0,015	0,210**	0,226**	0,293***	0,110	0,126
Habilidades adaptativas	0,090	0,206**	0,228***	0,137*	0,271***	0,168*	0,218***	0,234***	0,241***	0,257***	0,201**	0,255***	0,240***	0,189*	0.077
WHO-Motor			*n* = 152 [g]												
Motricidad gruesa	0,224**	0,126	0,282***	0,061	**0,703*** ** [b]										

* p<0,05, ** p<0,01, *** p<0,001. Correlaciones de Pearson en puntuaciones estandarizadas internamente; los errores estándar (EE) se calcularon utilizando métodos *bootstrap* , estratificando por grupo etario y sector socioeconómico (*n* =1.000 repeticiones). Las correlaciones de las escalas que miden el mismo dominio del desarrollo están resaltadas en negrita.

[a] Concurrencia mayor a la del ASQ-3, dominios coincidentes; [b] Concurrncia mayor a la del BDI-2, dominios coincidentes; [c] Concurrencia mayor a la del Denver-II, dominios coincidentes; [d] Concurrencia mayor a las del SFII, dominios coincidentes; [e] Concurrencia mayor a la concurrencia de la escala de motricidad fina-adaptativa con la escala cognitiva del Bayley-III. [f] Niños de 8 a 18 meses. [g] Niños de 6 a 15 meses.

Tabla 7: Correlaciones entre las escalas del Bayley-III y las pruebas cortas, para el 25% más pobre y el 25% más rico de la muestra, niños de 6 a 42 meses

	Bayley-III, 25% más pobre					Bayley-III, 25% más rico				
	Cognitiva	**Lenguaje receptivo**	**Lenguaje expresivo**	**Motricidad fina**	**Motricidad gruesa**	**Cognitiva**	**Lenguaje receptivo**	**Lenguaje expresivo**	**Motricidad fina**	**Motricidad gruesa**
ASQ-3			*n* = 186					*n* = 151		
Resolución de problemas	**0,106**	0,136	0,248***	0,069	0,018	**0,176***	0,236**	0,168*	0,118	0,104
Comunicación	0,254***	**0,224****	**0,465*****	0,244***	0,154*	0,166*	**0,253****	**0,403*****	0,043	0,149
Motricidad fina	0,153*	0,109	0,230**	**0,203****	0,205**	0,123	0,098	0,150	**0,251****	0,127
Motricidad gruesa	0,108	0,055	0,026	0,152*	**0,279*****	0,080	0,102	0,025	-0,029	**0,396*****
Personal-social	0,073	0,034	0,148*	0,070	0,048	0,099	0,144	0,186*	0,198*	0,140
Denver-II			*n* = 184					*n* = 146		
Lenguaje	0,329***	**0,351*****	**0,510*****	0,262***	0,274***	0,335***	**0,434*****	**0,587*****	0,220**	0,281***
Motricidad fina-adaptativa	**0,381*****	0,350***	0,292***	**0,325*****	0,356***	**0,301*****	0,303***	0,242**	**0,362*****	0,105
Motricidad gruesa	0,338***	0,349***	0,234**	0,310***	**0,560*****	0,108	0,172*	0,160	0,264**	**0,483*****
Personal-social	0,193**	0,132	0,110	0,243***	0,178*	0,181*	0,196*	0,266**	0,199*	0,203*
SFI & SFII (MacArthur)			*n* = 124					*n* = 97		
Lenguaje receptivo[a]	0,177	**0,230**	**0,149**	0,143	0,145	0,079	**0,194**	**0,342***	-0,094	0,158
Lenguaje expresivo[b]	0,241**	**0,261****	**0,441*****	0,138	0,251**	0,179	**0,390*****	**0,509*****	0,219*	0,176
BDI-2 (Battelle)			*n* = 148					*n* = 167		
Cognitiva	**0,279*****	0,372***	0,385***	0,244**	0,226**	**0,448*****	0,401***	0,349***	0,364***	0,261***
Comunicación	0,281***	**0,335*****	**0,511*****	0,232**	0,211*	0,381***	**0,385*****	**0,480*****	0,272***	0,263***
Motricidad	0,195*	0,255**	0,305***	**0,353*****	**0,344*****	0,242**	0,236**	0,150	**0,323*****	**0,333*****
Personal-social	-0,035	0,058	0,194*	-0,114	0,004	0,225**	0,227**	0,252***	0,078	0,107
Habilidades adaptativas	0,068	0,114	0,187*	0,175*	0,156	0,225**	0,255***	0,238**	0,283***	0,219**
WHO-Motor			*n* = 35					*n* = 36		
Motricidad gruesa[c]	0,105	-0,037	-0,123	0,029	**0,675*****	0,087	0,095	0,353*	0,096	**0,443****

* p<0,05, ** p<0,01, *** p<0,001. Correlaciones de Pearson en puntuaciones estandarizadas internamente; los errores estándar (EE) se calcularon utilizando métodos *bootstrap*, estratificando por grupo etario y sector socioeconómico (*n* =1.000 repeticiones). Las correlaciones de las escalas que miden el mismo dominio del desarrollo están resaltadas en negrita. [a] Niños de 8 a 18 meses. [b] Niños de 8 a 30 meses. [c] Niños de 6 a 15 meses.

APÉNDICE

Apéndice I: Sitios web de las editoriales[21], modificaciones lingüísticas y otras adaptaciones realizadas en los ítems de las pruebas

Bayley-III
Editorial: Pearson
http://www.pearsonclinical.com/childhood/products/100000123/bayley-scales-of-infant-and-toddler-development-third-edition-bayley-iii.html#tab-pricing

El Bayley-III estuvo disponible en español por primera vez a mediados de 2015, por lo que tuvimos que traducir la versión en inglés de la prueba al español de Colombia, y luego traducirla al inglés de nuevo.

Además, hubo que modificar las siguientes imágenes en las escalas de lenguaje:
- En varios ítems, para la acción 'lavando', cambiamos la imagen de una lavadora que aparece en el cuaderno de estímulos por un lavadero normal. Mantuvimos el mismo verbo.
- En varios ítems, reemplazamos el verbo 'aspirar' por 'barrer', así como también modificamos la imagen para que fuera acorde.

ASQ-3
Editorial: Brookes Publishing Co.
Paquete de materiales: http://products.brookespublishing.com/ASQ-3-in-Spanish-Starter-Kit-P575.aspx
Paquete de manipulativos: http://products.brookespublishing.com/Ages-Stages-Questionnaires-Third-Edition-ASQ-3-Materials-Kit-P585.aspx

Modificamos las siguientes palabras de la versión original del ASQ-3 en español:
- Escala de motricidad gruesa. Ítems 9 (cuestionario 33) y 8 (cuestionario 36): Reemplazamos 'resbaladilla' por 'rodadero', una palabra utilizada con mayor frecuencia en Colombia.
- Escala de motricidad fina. Ítems 3, 5, 7 (cuestionario 6), ítems 1, 3, 5, 8 (cuestionario 8), ítems 2, 5, 7 (cuestionarios 9 y 10), ítems 2, 4 (cuestionario 12) e ítem 1 (cuestionario 14): Reemplazamos 'cheerio' por 'bolita de cereal'.
- Escala de resolución de problemas. Ítem 8 (cuestionario 8), ítem 5 (cuestionarios 9 y 10), ítems 2, 7 (cuestionario 12), ítems 4, 8 (cuestionario 14), ítems 2, 6 (cuestionario 16), ítems 3, 6 (cuestionario 18), ítem 8 (cuestionario 20), ítem 5 (cuestionario 22) e ítem 2 (cuestionario 24): Reemplazamos 'cheerio' por 'bolita de cereal'.
- Escala de comunicación. Ítems 9 (cuestionario 24) y 6 (cuestionario 27): A fin de que las instrucciones tuvieran más sentido, modificamos "pon el zapato encima de la mesa y pon el libro debajo de la silla" por "pon el libro sobre la mesa y pon el zapato debajo de la silla".

Denver-II
Editorial: Denver Developmental Materials Inc.

[21] Todos los sitios web fueron consultados por última vez el 25 de marzo de 2016.

http://denverii.com/
Fue necesario traducir algunas partes del manual de administración. Además, modificamos los siguientes ítems de las hojas de respuesta y las instrucciones de las pruebas.

- Ítems 1 y 7. Corregimos faltas ortográficas.
- Ítems 9 y 11. Añadimos la palabra 'dedos' para que las instrucciones fueran más claras.
- Ítems 21 y 24. Modificamos las instrucciones de las hojas de respuesta para que coincidieran con las del manual.
- Ítem 25. Reemplazamos 'banana' por 'banano' y 'cerca' por 'reja', dado que estas palabras se utilizan comúnmente en Colombia.
- Ítem 26. Corregimos faltas ortográficas y las instrucciones de la hoja de respuesta para que coincidieran con las del manual.

SFI
Editorial: Brookes Publishing Co.
http://www.brookespublishing.com/resource-center/screening-and-assessment/cdi/

Consejo Consultivo del CDI
http://mb-cdi.stanford.edu/board.html

La versión original fue desarrollada para México. A fin de garantizar la comprensión y equivalencia lingüística de todos los ítems de las pruebas para Colombia, realizamos las siguientes modificaciones:

- Reemplazamos 'guagua' por 'guaguau'
- Reemplazamos 'camión/troca' por 'bus'
- Quitamos la palabra 'coche' y la incluimos como una opción dentro de 'carro/coche'
- Reemplazamos 'tortilla' por 'arepa'
- Reemplazamos 'botella/mamilla' por 'tetero'
- Añadimos la palabra 'plata' como alternativa a 'dinero' ('dinero/plata')
- Reemplazamos 'lavabo' por 'lavamanos'
- Quitamos la palabra 'templo' y la incluimos como una opción dentro de 'templo/iglesia'
- Reemplazamos 'byebye' por 'chao'

SFII
Editorial: Brookes Publishing Co.
http://www.brookespublishing.com/resource-center/screening-and-assessment/cdi/

De modo similar, y además de modificar las palabras 'guagua', 'carro/coche', 'camión/troca', 'botella/mamilla', 'iglesia/templo' y 'byebye' (véase modificaciones a SFI), realizamos los siguientes cambios a fin de garantizar un equivalente lingüístico de todos los ítems de las pruebas para Colombia:

- Reemplazamos 'víbora' por 'culebra/serpiente'
- Reemplazamos 'plátano/banana' por 'plátano/banano'
- Reemplazamos 'calabaza' por 'tomate'
- Reemplazamos 'chícharo' por 'pollo'
- Reemplazamos 'cerillos' por 'fósforos'

BDI-2

Editorial: Riverside Publishing
https://secure.riversidepublishing.com/products/bdi2/pricing.html

Incluso contando con la versión en español, hubo que traducir del inglés al español los cuadernos de aplicación (manuales), en los que se especifican las instrucciones para administrar los ítems y puntuarlos de manera correcta. Asimismo, también se tradujo el texto que aparecía en el libro de imágenes (libro de cuentos).

Además, hubo que modificar los ítems traducidos del siguiente modo para asegurar su comprensión en toda la muestra de estudio:

- Ítem 2: Cambiamos 'mama' por 'alimentarse'.
- Ítem 3: Reemplazamos 'traga' por 'pasa la comida'.
- Ítem 37: Reemplazamos 'centavos' por 'pesos'.
- Ítem 80: Cambiamos 'cordel' por 'pita'.
- Ítems 10, 25, 26, 31, 332 y 94: modificamos las instrucciones, ya que su comprensión resultaba difícil para los padres y niños.
- Ítem 60: realizamos cambios en las instrucciones debido a que estas estaban basadas en ilustraciones acompañadas de palabras en inglés.
- Ítem 95: reemplazamos el dibujo del 'tren' por el dibujo de un 'bus'.

WHO-Motor
Editorial: WHO
http://www.who.int/childgrowth/standards/motor_milestones/en/

Tradujimos las instrucciones para la administración y puntuación de la prueba del inglés al español de Colombia.

Apéndice II: Estandarización interna de las puntuaciones utilizando promedios y desviaciones estándar condicionados a la edad.

Para cada escala, eliminamos los efectos del evaluador del puntaje bruto mediante una regresión de los puntajes brutos sobre indicadores (dummies) para cada evaluador utilizando el método de Mínimos Cuadros Ordinarios (MCO). Construimos los valores residuales de estas regresiones y los estandarizamos por edad utilizando métodos no paramétricos del siguiente modo: En primer lugar, computamos el promedio condicionado a la edad utilizando los valores predichos (estimados) de la regresión en (1), que estimamos mediante el método *kernel-weighted local polynomial smoothing*:

$$\mathrm{Y_i} = \mathrm{f}(Xi\ \ ) + \varepsilon_\mathrm{i} \qquad \forall\ i \tag{1}$$

donde $\mathrm{Y_i}$ es el valor residual del puntaje bruto del niño *i* en una escala determinada de una regresión sobre indicadores (dummies) para cada evaluador. X_i es la edad del niño en días. Luego, estimamos una regresión de los cuadrados de los residuos en (1) sobre la edad del niño (expresada en días), tal y como muestra la regresión *kernel-weighted local polynomial* en (2):

$$(\mathrm{Y_i} - \hat{\mathrm{f}}_i)^2 = \mathrm{g}(X_i) + v_\mathrm{i} \qquad \forall\ i \tag{2}$$

La estimación de la desviación estándar condicionada a la edad corresponde a la raíz cuadrada de los valores ajustados (predichos) $\hat{\mathrm{g}}_i$ en (2). Por último, calculamos el puntaje z ajustado por edad internamente, $\mathrm{ZY_i}$, sustrayendo del residuo del puntaje bruto el promedio condicionado a la edad en la muestra estimado en (1) y dividiendo por la desviación estándar condicionada a la edad en la muestra obtenida mediante la regresión en (2). Más concretamente:

$$\mathrm{ZY_i} = \frac{\mathrm{Y_i} - \hat{\hat{\mathrm{f}}}_i}{\sqrt{\hat{\mathrm{g}}_i}} \qquad \forall\ i \tag{3}$$

Como resultado se obtuvieron puntuaciones estandarizadas internamente con una distribución normal, con promedio cero en la totalidad del rango etario (distribuciones disponibles bajo solicitud).

Apéndice III: Tablas Apéndice

Tabla A1: Puntajes brutos y puntuaciones compuestas del Bayley-III y puntajes brutos de las pruebas cortas—para la totalidad de la muestra y por edad

	Promedio	DE	Promedio	DE	Promedio	DE	Promedio	DE
Puntajes brutos del Bayley-III	(*n* = 1.311)		(*n* = 443)		(*n* = 454)		(*n* = 414)	
Cognitiva	58,74	14,06	42,38	8,04	61,59	5,81	73,13	4,07
Lenguaje receptivo	25,44	9,12	15,27	3,46	26,36	5,14	35,32	3,49
Lenguaje expresivo	25,25	10,27	14,30	4,03	25,64	5,04	36,53	5,71
Motricidad fina	39,32	10,07	28,44	5,30	39,63	3,95	50,61	4,45
Motricidad gruesa	52,38	11,57	39,10	9,02	55,62	3,66	63,05	2,76
Puntuaciones compuestas del Bayley-III								
Cognitiva	98,37	8,91	103,86	9,44	95,73	7,93	95,40	6,25
Lenguaje	96,49	9,88	99,27	9,95	93,30	10,14	97,02	8,45
Motricidad	99,55	10,85	95,79	11,55	99,47	10,08	103,66	9,33
ASQ-3 (9 ítems)	(*n* = 664)		(*n* = 221)		(*n* = 224)		(*n* = 219)	
Resolución de problemas	46,86	15,27	46,77	12,93	46,59	15,93	47,24	16,75
Comunicación	46,73	18,13	44,41	14,34	42,99	17,76	52,90	20,27
Motricidad fina	46,61	15,19	45,68	15,38	44,49	13,14	49,73	16,47
Motricidad gruesa	50,14	17,52	44,93	20,56	51,81	14,20	53,68	16,04
Personal-social	48,06	14,46	47,87	14,91	44,98	11,93	51,39	15,67
Denver-II	(*n* = 658)		(*n* = 225)		(*n* = 221)		(*n* = 212)	
Lenguaje	20,41	6,32	13,86	2,82	20,83	3,29	26,93	4,01
Motricidad fina-adaptativa	18,77	4,39	13,81	2,60	19,73	1,91	23,03	1,89
Motricidad gruesa	20,90	5,48	14,43	3,49	22,69	1,99	25,89	1,83
Personal-social	15,93	5,04	10,09	3,23	17,35	1,75	20,65	1,97
SFI (MacArthur)[a]	(*n* = 192)		(*n* = 192)					
Lenguaje receptivo	44,56	20,85	44,56	20,85				
Lenguaje expresivo	6,39	6,98	6,39	6,98				
SFII (MacArthur)[b]	(*n* = 226)				(*n* = 226)			
Lenguaje expresivo	52,44	26,18			52,44	26,18		
BDI-2 (Battelle)	(*n* = 635)		(*n* = 215)		(*n* = 227)		(*n* = 193)	
Cognitiva	16,93	4,01	13,27	2,66	17,03	1,70	20,89	3,30
Comunicación	17,40	6,57	10,54	4,09	18,16	2,63	24,15	4,09
Motricidad	18,52	6,38	11,21	4,49	20,36	2,52	24,51	2,15
Personal-social	18,76	5,79	13,32	3,85	19,54	3,41	23,89	4,47
Habilidades adaptativas	16,39	5,51	10,46	3,22	17,77	3,03	21,39	3,32
WHO-Motor[c]	(*n* = 152)		(*n* = 152)					
Motricidad gruesa	3,99	2,08	3,99	2,08				

[a] Niños de 8 a 18 meses. [b] Niños de 19 a 30 meses. [c] Niños de 6 a 15 meses.

Tabla A2: Correlaciones entre las pruebas cortas y las puntuaciones compuestas del Bayley-III, por grupo etario

	Bayley-III, 6-18 meses			Bayley-III, 19-30 meses			Bayley-III, 31-42 meses		
	Cognitiva	Lenguaje	Motricidad	Cognitiva	Lenguaje	Motricidad	Cognitiva	Lenguaje	Motricidad
ASQ-3		*n* = 221			*n* = 224			*n* = 219	
Resolución de problemas	**0,161***	0,089	0,024	**0,063**	0,130	0,081	**0,307*****	0,441***	0,236***
Comunicación	0,143*	**0,230*****	0,191**	0,168*	**0,404*****	0,076	0,384***	**0,547*****	0,321***
Motricidad fina	0,024	0,035	**0,067**	0,109	0,155*	**0,184****	0,308***	0,299***	**0,401*****
Motricidad gruesa	0,095	0,038	**0,368*****	-0,072	0,013	**0,109**	0,186**	0,084	**0,215****
Personal-social	0,195**	0,188**	0,153*	0,000	0,093	0,089	0,060	0,198**	0,139*
Denver-II		*n* = 225			*n* = 221			*n* = 212	
Lenguaje	0,051	**0,191****	0,054	0,228***	**0,536*****	0,148*	0,490***	**0,616*****	0,406***
Motricidad fina-adaptativa	**0,166***	0,083	**0,139***	**0,344*****	0,367***	**0,314*****	**0,387*****	0,387***	**0,456*****
Motricidad gruesa	0,092	0,016	**0,372*****	0,005	0,150*	**0,274*****	0,196**	0,210**	**0,323*****
Personal-social	0,196**	0,159*	0,137*	0,054	0,153*	0,210**	0,231***	0,158*	0,112
SFI & SFII (MacArthur)		*n* = 192[a]			*n* = 226				
Lenguaje receptivo	0,140	**0,181***	0,103						
Lenguaje expresivo	0,035	**0,178***	0,029	0,179**	**0,472*****	0,116			
BDI-2 (Battelle)		*n* = 215			*n* = 227			*n* = 193	
Cognitiva	**0,253*****	0,173*	0,273***	**0,205****	0,307***	0,180**	**0,490*****	0,459***	0,324***
Comunicación	0,133	**0,065**	0,228***	0,302***	**0,513*****	0,192**	0,422***	**0,630*****	0,332***
Motricidad	0,077	0,018	**0,298*****	0,218***	0,237***	**0,331*****	0,355***	0,309***	**0,388*****
Personal-social	0,058	0,120	0,101	0,157*	0,295***	-0,046	0,127	0,226**	0,051
Habilidades adaptativas	0,031	0,189**	0,180**	0,066	0,186**	0,198**	0,220**	0,286***	0,171*
WHO-Motor		*n* = 152[b]							
Motricidad gruesa	0,190*	0,149	**0,513*****						

* p<0,05, ** p<0,01, *** p<0,001. Correlaciones de Pearson entre las puntuaciones compuestas del Bayley-III y las puntuaciones estandarizadas internamente de las pruebas cortas. Los errores estándar (EE) se calcularon mediante métodos *bootstrap* , estratificando por edad y sector socioeconómico (*n* =1.000 repeticiones). Las correlaciones de las escalas que miden el mismo dominio del desarrollo están resaltadas en negrita.

[a] Niños de 8 a 18 meses. [b] Niños de 6 a 15 meses.