Rubio-Codina, Marta; Grantham-McGregor, Sally

Working Paper

Validez predictiva de pruebas cortas comúnmente usadas para medir el desarrollo infantil en estudios a gran escala

IDB Working Paper Series, No. IDB-WP-1174

Provided in Cooperation with:
Inter-American Development Bank (IDB), Washington, DC

Suggested Citation: Rubio-Codina, Marta; Grantham-McGregor, Sally (2020) : Validez predictiva de pruebas cortas comúnmente usadas para medir el desarrollo infantil en estudios a gran escala, IDB Working Paper Series, No. IDB-WP-1174, Inter-American Development Bank (IDB), Washington, DC,
https://doi.org/10.18235/0002883

This Version is available at:
https://hdl.handle.net/10419/237469

Banco Interamericano de Desarrollo

DOCUMENTO DE TRABAJO DEL BID Nº IDB-WP-1174

Validez predictiva de pruebas cortas comúnmente usadas para medir el desarrollo infantil en estudios a gran escala

Marta Rubio-Codina
Sally Grantham-McGregor

Banco Interamericano de Desarrollo
División de Protección Social y Salud

Noviembre 2020

BID
Banco Interamericano de Desarrollo

Validez predictiva de pruebas cortas comúnmente usadas para medir el desarrollo infantil en estudios a gran escala

Marta Rubio-Codina
Sally Grantham-McGregor

Noviembre 2020

Catalogación en la fuente proporcionada por la
Biblioteca Felipe Herrera del
Banco Interamericano de Desarrollo
Rubio-Codina, Marta.
Validez predictiva de pruebas cortas comúnmente usadas para medir el desarrollo infantil en estudios a gran escala / Marta Rubio-Codina, Sally Grantham-McGregor.
p. cm. — (Documento de trabajo del BID ; 1174)
Incluye referencias bibliográficas.
1. Child development-Colombia-Testing. 2. Child development-Colombia-Forecasting.
3. Psychological tests for children-Colombia. I. Grantham-McGregor, Sally M. II.
Banco Interamericano de Desarrollo. División de Protección Social y Salud. III. Título.
IV. Serie.
IDB-WP-1174

http://www.iadb.org

scl-sph@iadb.org

www.iadb.org/SocialProtection

Validez predictiva de pruebas cortas comúnmente usadas para medir el desarrollo infantil en estudios a gran escala

Marta Rubio-Codina

Sally Grantham-McGregor

Banco Interamericano de Desarrollo

2020

Validez predictiva de pruebas cortas comúnmente usadas para medir el desarrollo infantil en estudios a gran escala[1]

Marta Rubio-Codina[2] Sally Grantham-McGregor[3]

Resumen*

Existe un creciente compromiso a nivel mundial para llevar a cabo intervenciones en la primera infancia con el objetivo de promover el desarrollo de millones de niños en situación de desventaja en países de ingresos bajos y medios que no logran alcanzar su pleno potencial de desarrollo. No obstante, los progresos en esta área se ven obstaculizados por la falta de pruebas de desarrollo factibles de ser usadas a escala. Es por esto por lo que contar con este tipo de pruebas es una necesidad de carácter urgente. Si bien se suelen utilizar pruebas de tamizaje o pruebas de un solo dominio ("pruebas cortas") como alternativas, se desconoce su validez predictiva en estas circunstancias. En 2011, se inició un estudio longitudinal en Colombia en el que psicólogos administraron las Escalas Bayley de Desarrollo Infantil (Bayley-III) a 1311 niños y niñas de entre 6 y 42 meses, a quienes también se les administró de forma aleatoria una de dos baterías de pruebas cortas en condiciones similares a las de una encuesta. Este estudio informó sobre la validez concurrente de las pruebas cortas comparadas con el Bayley-III ("prueba de referencia"). En 2016, a 940 niños de esa muestra, quienes en ese entonces tenían entre 6 y 8 años, se les administraron pruebas para medir el CI (Coeficiente Intelectual mediante la Escala de Inteligencia de Wechsler para Niños, WISC-V) y el desempeño académico (aritmética, comprensión lectora y vocabulario). Se comparó la capacidad para predecir el CI y el desempeño académico en la infancia intermedia entre las pruebas cortas, los Indicadores de Cuidado Familiar (FCI, por sus siglas en inglés), la talla para la edad, la desnutrición crónica (talla para la edad de -2 DE por debajo de la mediana) y el Bayley-III. La validez predictiva aumentó con la edad en todas las pruebas, y las escalas de desarrollo cognitivo y de lenguaje presentaron los puntajes más altos en la mayoría de los casos. La capacidad predictiva de todas las pruebas fue negligible en niños de 6 a 18 meses. A partir de esta edad, si bien el Bayley-III tuvo la mayor validez predictiva, la Prueba de Tamizaje del Desarrollo de Denver demostró ser la prueba corta más factible y con mayor validez. Esta prueba podría utilizarse con un menor riesgo de pérdida de validez en comparación con el Bayley-III. Por otra parte, los Inventarios MacArthur-Bates del Desarrollo de Habilidades Comunicativas en niños de 19 a 30 meses, al igual que el FCI en menores de 31 meses, predijeron el CI y el desempeño académico tan bien como el Bayley-III. Por último, el FCI presentó una validez predictiva más alta que la talla para la edad y el

[1] Este documento es una traducción, ligeramente modificada, del artículo original: Rubio-Codina M. and S. Grantham-McGregor. 2020. "Predictive validity in middle childhood of short tests of early childhood development used in large scale studies compared to the Bayley-III, the Family Care Indicators, height-for-age and stunting: A longitudinal study in Bogota, Colombia", *PLoS ONE,* 15(4): e0231317. https://doi.org/10.1371/journal.pone.0231317

[2] División de Protección Social y Salud, Banco Interamericano de Desarrollo, Washington, D.C. EE. UU.
[3] Facultad de Ciencias de la Salud Poblacional, Instituto de Salud Infantil de University College London, Londres, Reino Unido.

* Agradecemos a todas las familias participantes y a las escuelas primarias a las que asistían los niños en el estudio por prestarnos sus instalaciones para realizar las pruebas. Asimismo, extendemos nuestro agradecimiento a los evaluadores y encuestadores de ambas rondas; a Andrea Solano y Marlenny Escribano por llevar a cabo una excelente capacitación y coordinación de campo; a Paula Bernal por su invaluable asesoramiento técnico sobre las pruebas y sus adaptaciones; y a María Adelaida Martínez y Marta Dormal por su asistencia en la investigación. Este trabajo no hubiera sido posible sin el apoyo y aliento de María Caridad Araujo y Ana Lucía Muñoz.

indicador de desnutrición crónica, por lo que podría añadirse a este último para su uso como un indicador del desarrollo infantil a nivel poblacional.

Palabras claves: evaluación del desarrollo, prueba de diagnóstico, validez predictiva, desarrollo cognitivo, lenguaje, desarrollo motor, bebés y niños pequeños, FCI, desnutrición crónica, talla para la edad, estudios a gran escala, países de ingresos bajos y medios.

Códigos JEL: J1, I1, I2, I3

Contenido

1. Introducción

En los países de ingresos bajos y medios, alrededor de 250 millones de niños y niñas menores de 5 años no logran alcanzar su potencial de desarrollo (Black et al. 2017). Reconociendo que en las primeras etapas de la vida se sientan las bases de la salud y el bienestar para la edad adulta, existe un creciente compromiso a nivel mundial para implementar intervenciones destinadas a promover el desarrollo infantil temprano (DIT) a escala con el objetivo de abordar este problema y fomentar el desarrollo de niños y niñas en situación de desventaja (Black et al. 2017). No obstante, el progreso de estos esfuerzos se ve obstaculizado por la falta de instrumentos de medición del desarrollo infantil confiables y válidos, que permitan recolectar datos con facilidad, en especial en niños menores de 3 años (Black et al. 2017; Fernald et al. 2017; McCoy et al. 2016). Estos instrumentos son esenciales tanto para monitorear y evaluar la efectividad de las intervenciones que se diseñen e implementen, como para medir los niveles de DIT a nivel poblacional.

En la actualidad, se están llevando a cabo iniciativas para desarrollar tanto instrumentos de medición poblacionales que sean universales (esto es, culturalmente neutros o muy fáciles de adaptar), válidos, de acceso gratuito y factibles de implementar a escala, como instrumentos a nivel individual adecuados para evaluar intervenciones —ambos para niños de 0 a 3 años (Cavallera et al. 2019; McCoy et al. 2016). Si bien las pruebas existentes que evalúan el desarrollo de manera integral, tales como las Escalas de Bayley (Bayley 1969, 2006), son sensibles a los impactos de intervenciones que promueven el DIT (Attanasio et al. 2014; Fernald et al. 2017; Hamadani et al. 2006; Nahar et al. 2009), de acuerdo con nuestra experiencia y la de muchos otros investigadores, su administración demanda una gran cantidad de tiempo y recursos, y requiere de evaluadores altamente capacitados que posean un cierto nivel de especialización técnica (Fernald et al. 2017). Cabe destacar que estos aspectos dificultan su aplicación a escala (Rubio-Codina, Araujo, Attanasio, Muñoz, et al. 2016). En consecuencia, a modo de alternativa, cada vez es más frecuente el uso de pruebas de tamizaje (diseñadas para identificar a niños en riesgo de presentar retrasos en el desarrollo) o pruebas de un solo dominio (como por ejemplo, el dominio de lenguaje) ya existentes, tanto para encuestas poblacionales (Fernald et al. 2012) como para evaluaciones de programas (Fernald y Hidrobo 2011; Macours, Schady, y Vakis 2012), puesto que son más fáciles, rápidas y baratas de administrar, y dado que ya están disponibles para uso. No obstante, se desconoce la confiabilidad y validez de estas pruebas cuando se utilizan a escala para detectar diferencias en los niveles del desarrollo dentro del rango normal de habilidades o para monitorear los avances en materia de desarrollo a nivel poblacional, en lugar de usarse para el fin para el que fueron diseñadas —esto es, para el tamizaje de niños en alto riesgo de presentar retrasos en el desarrollo. Por lo tanto, es necesario determinar estas propiedades (Fernald et al. 2017). En términos generales, es imperativo identificar —de entre los ya disponibles— instrumentos de medición del DIT confiables, válidos y factibles para uso en muestras de gran tamaño, hasta que se puedan utilizar los instrumentos a nivel poblacional e individual que están actualmente en desarrollo.

En un estudio anterior, evaluamos la confiabilidad, factibilidad de uso y validez concurrente en relación con las Escalas Bayley de Desarrollo Infantil (*Bayley Scales of Infant and Toddler Development*, tercera edición, Bayley-III) (Bayley 2006) de cinco pruebas que se utilizan comúnmente en evaluaciones y estudios a escala, en niños de 6 a 42 meses en Bogotá, Colombia (Rubio-Codina, Araujo, Attanasio, Muñoz, et al. 2016). Las pruebas (de aquí en adelante, "pruebas cortas") incluyeron 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, 1992), el Inventario del Desarrollo de Battelle (*Battelle Developmental Inventory*, segunda edición, BDI-2) (Newborg 2005)— y dos pruebas de un solo dominio —los Hitos del Desarrollo Motor Grueso de la Organización Mundial de la Salud (OMS) (*WHO-Gross Motor Milestones*, WHO-Motor) (WHO Multicentre Growth Reference Study Group 2006; Wijnhoven et al. 2004) y 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). Estas pruebas cortas se administraron en los hogares por encuestadores debidamente capacitados pero que no tenían experiencia específica ni conocimiento previo en temas de DIT. Por lo tanto, se administraron bajo condiciones factibles de implementar en estudios a escala. Por el contrario, el Bayley-III, que consideramos nuestro "estándar de oro" o prueba de referencia, fue aplicado por psicólogos en un centro a fin de minimizar las distracciones y estandarizar la administración tanto como posible. Por lo tanto, se llevó a cabo en mejores condiciones.

El Bayley-III fue la prueba más costosa y cuya administración demandó más tiempo. Si bien el BDI-2 también supuso costos elevados y demandó mucho tiempo, esta prueba duró 20 minutos menos que el Bayley-III, en promedio. La duración del Denver-II y el ASQ-3 fue de un tercio o menos, por lo que se encuentran en una posición intermedia tanto en términos de tiempo como de costos de administración. Las pruebas de un solo dominio fueron las más rápidas de administrar, dado que demoraron no más de 8 minutos, en promedio, y fueron relativamente menos costosas (el WHO-Motor, por ejemplo, se administró de forma gratuita). La validez concurrente de las escalas cognitivas, de lenguaje y de motricidad fina de dichas pruebas con las escalas correspondientes —es decir, aquellas que miden los mismos dominios— del Bayley-III aumentó con la edad: las correlaciones fueron bajas en menores de 19 meses, de bajas a moderadas en niños de 19 a 30 meses y de moderadas a altas en niños mayores de 30 meses. Mientras que el ASQ-3 exhibió un desempeño deficiente en niños menores de 31 meses, el SFII (que mide el lenguaje expresivo por reporte del cuidador) mostró una correlación relativamente alta con la escala de lenguaje del Bayley-III en menores de 30 meses. Combinando criterios de factibilidad y validez, el Denver-II demostró ser la mejor prueba corta multidimensional (Rubio-Codina, Araujo, Attanasio, Muñoz, et al. 2016).

Antes de elegir una prueba, es fundamental conocer su capacidad de predecir el desarrollo en el futuro (Snow y Hemel 2008). Por lo tanto, 5,5 años más tarde, revaluamos a los niños de la muestra, que en este entonces tenían entre 6 y 8 años, a fin de determinar la validez predictiva de las pruebas cortas —en comparación con el Bayley-III— del Coeficiente Intelectual (CI), evaluado en base a la quinta edición de la Escala Wechsler de Inteligencia para Niños (*Wechsler Intelligence Scale for Children*, WISC-V), y de un índice de desempeño académico que combina puntuaciones en aritmética, comprensión lectora y vocabulario. Asimismo, se analizó la capacidad predictiva del desarrollo futuro de la talla para la edad y del indicador de desnutrición crónica (talla para la edad de -2 desviaciones estándar [DE] por debajo de la mediana establecida por la OMS) en la primera infancia, puesto que esta última medición se ha utilizado en reiteradas ocasiones como un indicador aproximado (proxy) del desarrollo infantil global (Black et al. 2017; Grantham-McGregor et al. 2007). También se analizó la capacidad predictiva de los Indicadores de Cuidado Familiar (*Family Care*

Indicators, FCI), una medición de la calidad del ambiente en el hogar que se ha evaluado en un gran número de encuestas internacionales como un factor protector del DIT (Bornstein et al. 2012; Engle et al. 2011; MICS-ECDI s. f.). Al igual que con el análisis de validez concurrente reportado en Rubio-Codina, Araujo, Attanasio, Muñoz, et al. (2016), y debido al amplio rango de edad de los niños incluidos en la muestra (de 36 meses) así como al ritmo acelerado que caracteriza el desarrollo de los niños de 6 a 42 meses, el análisis de validez predictiva se realizó en intervalos de 12 meses de edad: de 6 a 18 meses, de 19 a 30 meses y de 31 a 42 meses. Estos intervalos corresponden a los grupos etarios más pequeños posible dados los tamaños de muestra disponibles por grupo.

Cabe destacar que la evidencia existente sobre la validez predictiva de las pruebas cortas para medir el DIT cuando se administran en encuestas de hogar a escala es muy limitada. Un estudio en zonas rurales de Bangladés (Hamadani et al. 2013), halló correlaciones bajas (r =[0,21-0,25]) pero significativas entre los reportes maternos mensuales acerca de la edad en la que el niño logra ciertos hitos de motricidad gruesa (caminar y pararse solo) y el CI a los 5 años. Del mismo modo, en el mismo contexto, Hamadani et al. (2010) encontraron que una prueba de lenguaje para niños de entre 12 y 18 meses, desarrollada localmente a partir de los Inventarios MacArthur-Bates de Desarrollo de Habilidades Comunicativas (Fenson et al. 2002) y administrada mediante reportes maternos, predijo de manera significativa el CI a los 5 años (r =[0,37-0,41]).

Partimos de las siguientes hipótesis de estudio: (i) la validez predictiva aumentaría con la edad de los niños, tal como se observó en el caso de la validez concurrente (Rubio-Codina, Araujo, Attanasio, Muñoz, et al. 2016); (ii) el Bayley-III presentaría la validez predictiva más alta, dado que es una prueba de diagnóstico; (iii) las escalas de desarrollo cognitivo y de lenguaje obtendrían un mejor desempeño que las escalas restantes, ya que las mediciones de CI están conformadas únicamente por funciones cognitivas y de lenguaje (Wechsler 2014); y (iv) el ambiente en el hogar (FCI), la talla para la edad y la desnutrición crónica presentarían una menor capacidad predictiva que las pruebas cortas, puesto que no evalúan dominios del desarrollo per se, aunque podrían utilizarse como proxies debido a su asociación con el desarrollo infantil (Black et al. 2017; Bornstein et al. 2012; Engle et al. 2011; Grantham-McGregor et al. 2007).

Este estudio no se diseñó ni cuenta con el poder suficiente para analizar la sensibilidad o especificidad de las pruebas de tamizaje para identificar a niños en alto riesgo de retraso en su desarrollo (Rubio-Codina, Araujo, Attanasio, Muñoz, et al. 2016). Además, se excluyó del estudio a niños con discapacidad o con puntuaciones menores a 70 en el Bayley-III. Por ello, nuestros resultados no tienen implicaciones relacionadas con el uso de las pruebas de tamizaje analizadas para detectar a niños que podrían encontrarse en riesgo y que requieren de evaluaciones adicionales. El objetivo de este estudio fue investigar la capacidad para predecir la inteligencia y el desempeño académico futuros de varias pruebas de tamizaje y de otras pruebas cortas e indicadores, utilizados con frecuencia para evaluar intervenciones o para medir el desarrollo infantil a nivel poblacional.

2. Materiales y métodos

2.1. Diseño y participantes en el estudio

Bogotá está dividida en seis sectores socioeconómicos, denominados estratos y que se definen con base en la ubicación y la calidad de las viviendas y de la infraestructura. El estudio se llevó a cabo en los tres estratos más pobres, los cuales comprenden hogares de ingresos bajos y medios, y representan el 85% de la población de la ciudad. En 2011, se incluyeron en la muestra 1533 niños de 6 a 42 meses que vivían en diferentes manzanas de estos estratos. Dentro de cada estrato, las manzanas se seleccionaron de forma aleatoria, ponderando por el porcentaje de mujeres en edad fértil. Una vez elegidas las manzanas, se realizaron visitas puerta a puerta con el objetivo de identificar a todos los niños y niñas de 6 a 42 meses y se seleccionó aleatoriamente una submuestra para inclusión en el estudio, que luego se estratificó por edad. Se computaron los tamaños de los estratos —originalmente, 4 sectores socioeconómicos y 4 grupos etarios (de 6 a 14, de 15 a 23, de 24 a 32 y de 33 a 42 meses de edad)— con el objetivo de detectar diferencias en el desarrollo infantil entre ellos. La muestra final incluyó a 12 niños pertenecientes a un cuarto sector socioeconómico (de ingresos medios), inicialmente incluido en el estudio, pero que luego se eliminó debido a la alta tasa de rechazo a participar entre los hogares en este sector. Se excluyó también a mellizos y a niños con discapacidades evidentes. En aquellos hogares en los que había más de un niño elegible, se incluyó solo a uno de ellos, seleccionado aleatoriamente. Rubio-Codina et al. (2015) brinda más información sobre la muestra.

A fin de incluir el mayor número de pruebas posible en el estudio y evitar, al mismo tiempo, que estas supusieran un proceso agotador para los niños y sus familias, se asignó a cada uno de los 1533 niños participantes a una de dos baterías de pruebas cortas, de manera aleatoria. Así, a aproximadamente la mitad de los niños se les administraron las pruebas cortas de la batería A, mientras que a la otra mitad se les administraron aquellas incluidas en la batería B. Entre 5 y 14 días después, se administró la prueba de Bayley-III en 1330 de estos niños. Se analizaron las puntuaciones de 1311 participantes. En 2016, realizamos un seguimiento y reevaluamos a tantos de estos niños como fue posible, quienes, en ese entonces tenían entre 6 y 8 años. El Gráfico 1 ofrece más información sobre el diseño del estudio y el flujo de participantes.

2.2. Procedimientos

Mediciones del desarrollo en la primera infancia. Las pruebas de la batería A incluyeron el ASQ-3 (Squires et al. 2009), el Denver-II (Frankenburg et al. 1990, 1992) y las listas de vocabulario del SFI y SFII (Jackson-Maldonado et al. 2003, 2012). Por otro lado, la batería B comprendía el BDI-2 (Newborg 2005) y el WHO-Motor (WHO Multicentre Growth Reference Study Group 2006; Wijnhoven et al. 2004).

Las pruebas de tamizaje multidimensionales —esto es, el ASQ-3, el Denver-II y el BDI-2— abarcan todo el rango etario. Estas tres pruebas incluyen los ítems relacionados con el lenguaje receptivo y expresivo dentro de una única escala de comunicación/lenguaje. Del mismo modo, la escala de motricidad del BDI-2 combina ítems de motricidad fina y gruesa; y la escala de motricidad fina-adaptativa del Denver-II incluye ítems cognitivos y de motricidad fina, mientras que la escala de resolución de problemas del ASQ-3 mide el desarrollo cognitivo. Tal como se explicó anteriormente (Rubio-Codina, Araujo, Attanasio, Muñoz, et al. 2016), todas las pruebas se administraron siguiendo las instrucciones de los manuales, a excepción del ASQ-3, cuya administración modificamos de la siguiente manera. Debido a los

bajos niveles de lectoescritura de algunos cuidadores, los ítems por reporte se completaron mediante entrevista a fin de garantizar que todas las madres entendieran las preguntas de manera similar. Asimismo, se evaluó al niño en aquellos casos en los que el cuidador no pudo dar una respuesta. Además, cuando el niño alcanzaba el techo del cuestionario apropiado para su edad, se evaluaban los siguientes tres ítems del cuestionario subsiguiente, que representaban un mayor grado de dificultad y excluyendo ítems coincidentes con los que ya se habían administrado en la prueba correspondiente a su edad. Esto disminuyó el porcentaje de niños en el techo de la prueba de un 10,5-15,5% a un 1,7-4,8%, según el dominio, lo que, en consecuencia, aumentó la variabilidad en las habilidades de desarrollo infantil medidas por el ASQ-3. Se han realizado adaptaciones del ASQ-3 similares a estas en otros estudios (Fernald et al. 2012). En la mayoría de los casos, los ítems del Denver-II y del BDI-2 se recolectaron mediante administración directa siguiendo los protocolos de los manuales de estas pruebas; a pesar de que hasta un 39% de los ítems del Denver-II (especialmente en las escalas de desarrollo personal-social y de lenguaje) se pueden administrar por reporte del cuidador y algunos ítems del BDI-2 también pueden obtenerse a través de estos reportes o mediante la observación del evaluador.

El WHO-Motor y los SF miden un solo dominio y cubren un rango de edad limitado. El WHO-Motor incluye seis hitos del desarrollo motor grueso para evaluar a niños de 6 a 18 meses de manera directa; no obstante, el análisis se limitó a niños de 6 a 15,9 meses debido a que la mayoría de los niños más grandes (91,9%) logró alcanzar todos los hitos. Las pruebas SFI y SFII miden lenguaje receptivo y expresivo (palabras que el niño "entiende" y palabras que el niño "entiende y dice") en niños de 8 a 18 meses, y lenguaje expresivo (palabras que el niño "dice") en niños de 19 a 30 meses, respectivamente, por reporte del cuidador.

Todas las pruebas cortas se administraron en los hogares y por encuestadores no especializados. Entre 5 y 14 días después de que se administraran las pruebas cortas, los psicólogos (de aquí en adelante, evaluadores), que desconocían los resultados que los niños habían obtenido en dichas pruebas, administraron el Bayley-III (Bayley 2006) a todos los niños en la muestra en un centro —por lo general, en la biblioteca de la red BibloRed más cercana al hogar del niño. Todos los dominios del Bayley-III se recolectaron mediante evaluación directa del niño, a excepción de la escala socioemocional para la cual se utilizaron reportes del cuidador principal. Asimismo, los evaluadores midieron la talla de los niños en el centro y siguiendo procedimientos estándar (World Health Organization 1983). Se estandarizaron los puntajes y se estimó la desnutrición crónica (talla para la edad de -2 DE por debajo de la mediana según lo establecido por la OMS) mediante el software WHO Anthro, 2011 de la OMS. Los evaluadores/encuestadores recibieron capacitación durante 6 semanas, incluyendo prácticas.

Previo al operativo de campo, el Bayley-III, el manual del BDI-2 y los manuales y hojas de respuesta del WHO-Motor se tradujeron al español y luego se realizó la traducción inversa. Dado que todas las pruebas cortas de la batería A estaban ya disponibles en español, no fue necesario traducirlas. Se realizaron pruebas piloto de todas las traducciones y versiones oficiales en español y, luego, se realizaron algunas modificaciones en la redacción y el estilo para reflejar mejor el español de Colombia. Del mismo modo, se ajustaron algunas imágenes al contexto local.

El Bayley-III fue la prueba más costosa (USD 1025 por kit; USD 4,89 por niño, cuando se realizó la evaluación) y su administración tuvo una duración de 83 minutos en promedio. El tiempo de administración del BDI-2 fue de 63 minutos y, al igual que el Bayley-III, demandó una gran cantidad de recursos (USD 405,70 por kit; USD 3,08 por niño); mientras que el Denver-II (27 minutos) y el ASQ-3 (20 minutos) se encuentran en una posición intermedia respecto del tiempo y el costo de administración (Denver-II: USD 200 por kit; USD 0,45 por niño; y el ASQ-3: USD 275 por kit, sin ningún valor adicional por niño). Tal como se había previsto, las pruebas de un solo dominio fueron las más rápidas (entre 6-y 8 minutos) y las menos costosas (SFI y SFII: USD 90 por kit incluyendo los dos formularios, USD 1 por niño; WHO-Motor: gratuita).

Las pruebas usadas en la primera infancia, así como sus costos, y los procedimientos de adaptación, capacitación y administración se describen en mayor detalle en publicaciones anteriores (Rubio-Codina, Araujo, Attanasio, y Grantham-McGregor 2016; Rubio-Codina, Araujo, Attanasio, Muñoz, et al. 2016).

Mediciones del desarrollo en la infancia intermedia. En 2016, se realizó un estudio de seguimiento y se volvió a evaluar a todos los niños que fue posible encontrar. Se midió su CI a través del WISC-V utilizando las 7 subpruebas que conforman la Escala Completa de Coeficiente Intelectual (*Full Scale Intelligence Quotient*, FSIQ): construcción con cubos, semejanzas, matrices, dígitos, claves, vocabulario y balanzas (Wechsler 2014). Se evaluó el desempeño académico a través de las subpruebas de aritmética (cálculos) y comprensión lectora incluidas en la tercera edición de la Prueba de Aprovechamiento de Woodcock-Muñoz (WM-III, Muñoz-Sandoval et al. 2005), la versión en español de la prueba *Woodcock-Johnson Test of Achievement* (Woodcock, McGrew, y Mather 2001) y un subconjunto de 75 palabras del Test de Vocabulario en Imágenes de Peabody (TVIP, Dunn et al. 1986), la versión en español del *Peabody Picture Vocabulary Test-Revised* (Dunn y Dunn 1981). Estas palabras se seleccionaron con base en información sobre niños que viven en zonas urbanas y de la misma edad y nivel socioeconómico que los niños de la muestra obtenidos de una encuesta longitudinal representativa a nivel nacional, la Encuesta Longitudinal Colombiana de la Universidad de Los Andes (ELCA 2010 & 2013). Primero se eligieron aquellas palabras correspondientes al rango de edad relevante que presentaban suficiente variabilidad y luego se las ordenó de acuerdo con el grado de dificultad. Las decisiones finales sobre qué palabras incluir y en qué orden se tomaron después del pilotaje. Estas acciones se llevaron a cabo a fin de simplificar la administración de la prueba y reducir el tiempo de evaluación.

Todas las pruebas de desarrollo para la infancia intermedia se administraron siguiendo las instrucciones de los manuales, a excepción del subconjunto de 75 palabras del TVIP, que se administraron por orden de dificultad hasta que el niño cometiera 3 errores consecutivos. Las pruebas se administraron de manera individual por psicólogos en las escuelas primarias de los niños (91,5%), y en algunas ocasiones, en otro centro (2,5%) o en los hogares de los participantes (6%). En ocasiones, las pruebas se llevaron a cabo en presencia de la madre, educador u otro familiar adulto (5,8%). El tiempo total de administración no superó los 90 minutos en ningún caso y además se incluyó un receso de entre 5 y 10 minutos a la mitad de cada evaluación.

Se tradujeron los manuales y hojas de respuesta del WISC-V, y se pilotearon dichas traducciones. El resto de materiales de las pruebas estaban disponibles en español con lo que no fue necesario traducirlos. No obstante, después del pilotaje, se realizaron algunas

modificaciones en la redacción y el estilo para reflejar de forma más adecuada el español de Colombia. No se consideró necesario realizar modificaciones adicionales.

Doce graduados en psicología recibieron capacitación durante cinco semanas y cada uno practicó la administración de cada prueba entre 15 y 20 veces, hasta que la confiabilidad intersujeto (entre evaluadores) alcanzara un nivel de concordancia de más de un 90% para los ítems de cada prueba. El capacitador observó el 2% de las evaluaciones del estudio y obtuvo una concordancia media mayor al 95% con los evaluadores (rango = [85-100%]). Cuando se consideró apropiado, se realizaron los comentarios y correcciones correspondientes respecto del desempeño del evaluador.

Encuesta de hogar. En ambas rondas, se visitaron los hogares de los niños para recopilar información sobre su composición y sobre otros aspectos socioeconómicos. En la primera ronda, se midió la calidad del ambiente en el hogar a través de los Indicadores de Cuidado Familiar (FCI, Kariger et al. 2012) de UNICEF, que incluyen los materiales y actividades de juego. Se les preguntó a los cuidadores sobre las actividades de juego que habían llevado a cabo los niños en compañía de un adulto durante la semana previa a la encuesta y se observó los tipos de juguetes con los que los niños jugaban usualmente. Para la evaluación de la calidad del ambiente del hogar en la infancia intermedia, se utilizó una adaptación de la prueba Observación para la Medición del Ambiente en el Hogar en la Infancia Intermedia (*Middle Childhood Home Observation for Measurement of the Environment*, MC-HOME) (Bradley et al. 1988; Caldwell y Bradley 2003).

Cuestiones éticas. Este estudio contó con la aprobación del comité de ética del Instituto de Ortopedia Infantil Roosevelt en Bogotá. Antes de cada prueba, los padres de los niños participantes firmaron un consentimiento informado.

3. Análisis estadístico

Para ambas rondas, los índices de riqueza respecto a la información sobre los activos y características de la vivienda se construyeron a partir de análisis de componentes principales con correlaciones policóricas (Rubio-Codina et al. 2015). Por otra parte, los puntajes totales para el FCI y el MC-HOME se obtuvieron sumando indicadores dicotómicos (0/1), con puntos de corte que variaban en función de la prevalencia empírica de la variable de interés para cada indicador del FCI. Es importante señalar que el puntaje del FCI incluye actividades de juego (leer/mirar libros ilustrados; contar historias; cantar canciones; llevar al niño fuera del hogar o salir a dar un paseo; jugar con distintos juguetes; hacer garabatos, dibujar o colorear; nombrar o contar cosas) y materiales de juego (juguetes para hacer o reproducir música; objetos para dibujar, escribir o pintar; libros para colorear; libros ilustrados; juguetes para realizar juegos de roles; juguetes para desplazarse o que requieren mucho movimiento físico; objetos para apilar o construir cosas; juguetes para aprender formas y colores).

La probabilidad de ser revaluado en 2016 se estimó mediante una regresión logística sobre las características iniciales (en la primera infancia), y el inverso de esta probabilidad se utilizó como un ponderador para ajustar por la pérdida de muestra en análisis de robustez posteriores. También analizamos las diferencias entre los participantes evaluados a quienes se les administraron las baterías A y B a través de pruebas *t* de diferencias en medias.

Las escalas de todas las pruebas se administraron y puntuaron de manera independiente, y los puntajes crudos (continuos) se construyeron siguiendo las instrucciones en los manuales. Dado que el Denver-II no cuenta con un puntaje crudo, añadimos los ítems con un puntaje de 1 a los ítems anteriores al nivel basal (de inicio), de acuerdo con los principios generales de cálculo de puntajes y lo realizado en estudios previos (Rubio-Codina, Araujo, Attanasio, Muñoz, et al. 2016). De modo similar, para el WHO-Motor, sumamos todos los ítems que el niño podía hacer a fin de obtener un puntaje crudo (Rubio-Codina, Araujo, Attanasio, Muñoz, et al. 2016).

Cabe señalar que ninguna de las pruebas contaba con normas establecidas para Colombia. Por lo tanto, estandarizamos los puntajes crudos internamente según la edad utilizando las medias y desviaciones estándar específicas para cada edad, calculadas de forma no paramétrica, luego de eliminar los efectos del evaluador/encuestador, tal como se había hecho en estudios anteriores (Rubio-Codina, Araujo, Attanasio, Muñoz, et al. 2016). Esto significa que, para cada valor de los residuos de los puntajes crudos sobre los efectos del evaluador o encuestador, construimos un puntaje z restando la media para cada edad y dividiendo por la desviación estándar específica para cada edad —ambas (media y desviación estándar) estimadas a partir de regresiones polinomiales locales (no paramétricas). A diferencia de cuando se utilizan normas establecidas en la población de referencia (es decir, puntajes estandarizados externamente) para cada prueba, este método de estandarización permite corregir por el efecto de la edad de una forma similar en todas las pruebas, lo cual facilita la comparación entre estas. Asimismo, este enfoque es menos sensible a valores extremos o tamaños de muestra limitados en comparación con otros métodos tradicionalmente implementados para estandarizar puntajes internamente y que habitualmente utilizan medias y desviaciones estándar para un intervalo determinado (p. ej., meses de edad) para calcular un puntaje z. El FSIQ se calculó a partir de la suma de los puntajes de las subpruebas del WISC-V estandarizados internamente y el puntaje de desempeño académico se construyó sumando los puntajes de aritmética, comprensión lectora y vocabulario también estandarizados internamente.

Analizamos la consistencia interna de las pruebas calculando el alfa de Cronbach (α); examinamos la confiabilidad test-retest a partir de correlaciones intraclase (ICC); y evaluamos las asociaciones entre las pruebas y las variables socioeconómicas utilizando correlaciones de Pearson (r), para todo el rango de edad y por grupos etarios de 12 meses. La capacidad para predecir el FSIQ y el desempeño académico de todas las pruebas cortas, el FCI, la talla para la edad y la desnutrición crónica se analizó por intervalos de 12 meses de edad, estimando las correlaciones de Pearson para cada intervalo. Se utilizaron puntajes estandarizados internamente en todas las correlaciones, lo cual equivale a calcular las correlaciones parciales controlando por los efectos del evaluador/encuestador y de la edad de manera flexible. En el caso del Denver-II, utilizamos la escala de motricidad fina adaptativa tanto para los análisis de cognición como de motricidad fina, puesto que existe evidencia de que esta escala incluye ítems relacionados con ambos dominios del desarrollo. En efecto, la escala cognitiva del Bayley-III contiene algunos ítems similares a los del Denver-II. Además, nuestro anterior análisis de validez concurrente (Rubio-Codina, Araujo, Attanasio, Muñoz, et al. 2016), mostró correlaciones algo superiores entre la escala de motricidad fina adaptativa del Denver-II y la escala de desarrollo cognitivo del Bayley-III, en comparación con las que se hallaron entre la primera escala mencionada y la escala de motricidad fina del Bayley-III en niños menores de 30 meses. Para los análisis de lenguaje receptivo y expresivo, utilizamos

las escalas de comunicación/lenguaje del ASQ-3, del Denver-II y del BDI-2, y empleamos la escala de desarrollo motor del BDI-2 tanto para evaluar la motricidad fina como la motricidad gruesa. De aquí en adelante, se hará referencia a las escalas que miden predominantemente el desarrollo cognitivo o del lenguaje mediante estos nombres.

Se utilizaron *P*-valores —calculados mediante métodos Bootstrap (Efron 1982) estratificando por los estratos establecidos en el diseño (grupo etario y sector socioeconómico)— para toda inferencia estadística y también para analizar si la validez predictiva de una prueba variaba significativamente entre los tres grupos etarios evaluados. Asimismo, estos *P*-valores calculados mediante el método Bootstrap se utilizaron para comparar la validez predictiva del Bayley-III, las pruebas cortas, el FCI, la talla para la edad y la desnutrición crónica.

Mientras que las pruebas dirigidas a niños pequeños suelen medir una variedad de dominios del desarrollo, la escala del FSIQ solo mide funciones cognitivas y del lenguaje. Por ello, para las pruebas cortas, en la descripción de resultados solo comparamos la validez predictiva entre las escalas de desarrollo cognitivo y de lenguaje; sin embargo, también presentamos los valores de validez predictiva del resto de escalas por su interés intrínseco y por completitud.

Para poder analizar con mayor profundidad el efecto de la edad en menores de 19 meses en relación con el Bayley-III, el FCI, la talla para la edad y la desnutrición crónica, repetimos el análisis en intervalos de 6 meses de edad en la primera infancia, dado que todas estas pruebas se encontraban disponibles para los niños en ambas baterías y por lo tanto había suficiente muestra por grupo etario de 6 meses. Para el resto de las pruebas, ya sean de la batería A o de la batería B, el tamaño de la muestra se consideró demasiado pequeño como para una nueva subdivisión.

Las correlaciones se clasificaron como muy bajas (r =0,10-0,19), bajas (r =0,20-0,39), moderadas (r =0,40-0,59) y altas (r =0,60-0,79) (Evans 1996).

Los análisis estadísticos se llevaron a cabo utilizando Stata 14.2 (StataCorp, College Station, TX).

4. Resultados

Se revaluó nuevamente a 940 niños (71,7%) de los 1311 niños iniciales que contaban con puntajes para el Bayley-III (obtenidos en la primera infancia). De estos 940 niños, tres presentaban puntuaciones menores a -3 DE de acuerdo con los puntajes del FSIQ estandarizados externamente y, por lo tanto, se excluyeron del análisis (Gráfico 1). Entre las principales razones que explican la pérdida de muestra se pueden mencionar la migración (50,4%), la incapacidad de contactar a la familia (20,7%), la negativa a seguir participando en el estudio (21,6%) y la negativa a que se les administre el WISC-V (7,3%). Se rastreó y realizó seguimiento a aquellas personas que habían migrado y se encontraban a una distancia de Bogotá de alrededor de una hora (en autobús). Se observó una mayor pérdida de muestra en relación con los hogares más pobres, las niñas y las madres más jóvenes o con un menor nivel educativo. Los niños evaluados en la infancia intermedia, a quienes inicialmente se les había administrado la batería A o la batería B, eran comparables en cuanto a las características que presentaban, a excepción de la edad de la madre (P =0,037) (Tabla 1),

que mostró una asociación significativa pero muy baja (*r* <0,127, *P* <0,001) con los resultados (no se muestra en este artículo).

Los puntajes crudos de todas las pruebas y subpruebas para la infancia intermedia, que se muestran en la Tabla A1 para todos los niños y por batería (A y B), aumentaron con la edad y el grado escolar (Tabla A2). Los puntajes promedio del WISC-V estandarizados externamente se encontraban por debajo de los de la muestra normativa (Tabla 1), y la consistencia interna era adecuada (*α*'s >0,6, excepto en dos casos) y se mantuvo constante a lo largo del tiempo (no se muestra en este artículo). La confiabilidad test-retest luego de 6 a 14 días también era adecuada ([ICC =0,39-0,88], ICC >0,6 para la mayoría de las pruebas). Tal como se esperaba con base en la teoría, el FSIQ y el desempeño académico mostraron estar asociados tanto con las características del hogar como entre sí (Tabla 2). Estas asociaciones eran superiores para niños de 7 y 8 años que para niños de 6 años (no se muestra).

En el Gráfico 2 se muestran las correlaciones de las pruebas cortas y el Bayley-III con el FSIQ, y en el Gráfico A1 se ilustran las correlaciones con el desempeño académico. En el panel superior de la Tabla 3 se presentan las correlaciones entre los puntajes iniciales (primera infancia) y los puntajes obtenidos en la infancia intermedia por dominio (o escala) y por grupo etario; y, en el panel inferior, se muestran los valores correspondientes al FCI, la talla para la edad y la desnutrición crónica. La Tabla 4 muestra las correlaciones que son significativamente diferentes entre sí para las escalas de lenguaje y de desarrollo cognitivo de las pruebas cortas y el Bayley-III, el FCI y la desnutrición crónica. En la Tabla A3 se muestran los tests de significancia estadística de las comparaciones de las correlaciones entre los grupos etarios. Las escalas de desarrollo cognitivo, de lenguaje y, en menor medida, la escala de motricidad fina, demostraron ser las más predictivas. En general, el Bayley-III presentó las correlaciones más altas, y en todas las pruebas, las correlaciones aumentaron con la edad en relación con la evaluación inicial, a excepción de la escala de desarrollo cognitivo del ASQ-3 donde las correlaciones disminuyeron en niños de 19 a 30 meses. El FSIQ y el desempeño académico mostraron una correlación alta entre sí (*r* =0,706, *P* <0,001, Tabla 2) y patrones de correlaciones similares también entre sí (Gráfico A1, Tabla 3). Dado esto, enfocamos la descripción de resultados en el FSIQ de ahora en adelante.

4.1. Pruebas multidimensionales

A pesar de que en niños menores de 19 meses la validez predictiva de las pruebas fue muy baja y casi que trivial (*r* <0,185 en todas), en algunos casos fue estadísticamente significativa (escalas de desarrollo cognitivo y de lenguaje receptivo del Bayley-III, y escala de desarrollo cognitivo del ASQ-3).

En niños de 19 a 30 meses, las escalas de desarrollo cognitivo y de lenguaje del Bayley-III, el Denver-II y el BDI-2 presentaron correlaciones bajas similares con el FSIQ pero significativas (*r* =0,330, *P* <0,001; *r* =0,229, *P* <0,01; *r* =0,221, *P* <0,01, para el desarrollo cognitivo respectivamente, y *r* =0,307, *P* <0,001; *r* =0,214, *P* <0,01; *r* =0,244, *P* <0,01, para el lenguaje o lenguaje expresivo). Las escalas de desarrollo cognitivo de estas tres pruebas obtuvieron correlaciones significativamente superiores a la del ASQ-3 (Tabla 4), la cual no presentó una asociación significativa con el FSIQ (Tabla 3). Asimismo, si bien la escala de lenguaje del ASQ-3 presentó correlaciones muy bajas (*r* =0,180, *P* <0,05), no difería significativamente del resto de las escales.

En niños de 31 a 42 meses, las correlaciones con el FSIQ que presentaban las cuatro pruebas multidimensionales aumentaron con respecto a las correlaciones observadas en niños más pequeños y fueron significativas. La validez predictiva de las escalas de desarrollo cognitivo (r =0,474, P <0,001) y de lenguaje receptivo (r =0,409, P <0,001) del Bayley-III, y de la escala de desarrollo cognitivo del Denver-II (r =0,422, P <0,001) mostró los valores más altos con niveles moderados, mientras que el resto de las escalas presentaron valores más bajos (r =0,271-0,386, P <0,05). Los resultados de la escala de desarrollo cognitivo del Bayley-III no fueron diferentes a los de las escalas de desarrollo cognitivo y de lenguaje del Denver-II, pero fueron significativamente superiores a los de las escalas de lenguaje del BDI-2 y del ASQ-3, y a los de la escala de desarrollo cognitivo del ASQ-3 (Tabla 4).

El aumento observado en las correlaciones entre los grupos etarios fue significativo para el grupo más joven (niños de 6 a 18 meses) y el grupo de mayor edad (niños de 30 a 41 meses) tanto para las escalas de desarrollo cognitivo como de lenguaje del Bayley-III y del Denver-II (Tabla A3). Este aumento también fue significativo para las escalas del Bayley-III en el grupo más joven y de edad intermedia (niños de 19 a 24 meses).

En general, las escalas de motricidad fina mostraron una capacidad predictiva más baja para el FSIQ que las escalas de desarrollo cognitivo y de lenguaje (r <0,353 para todas), a excepción de la del Denver-II que también incluye habilidades cognitivas. Las escalas de motricidad gruesa mostraron una correlación muy baja con el FSIQ (r <0,228 para todas), con los valores más altos para la escala motora del BDI-2, la cual combina ítems de motricidad gruesa con ítems de motricidad fina.

4.2. Pruebas de un solo dominio

Al igual que las otras pruebas cortas, el SFII no presentó una correlación significativa con el FSIQ en niños menores de 19 meses. Sin embargo, para el grupo etario de 19 a 30 meses, esta correlación (r =0,301, P <0,001) presentó valores bajos, similares a los de las escalas de desarrollo cognitivo y de lenguaje del Bayley-III y ligeramente superiores a los de las otras pruebas, aunque solo significativamente más altos que los de la escala de desarrollo cognitivo del ASQ-3 (P <0,000, Tabla 4). Por otro lado, el aumento en la validez predictiva del SFI (niños de 6 a 18 meses) y del SFII (niños de 19 a 24 meses) fue estadísticamente significativo (Tabla A3). Por último, el WHO-Motor no predijo de manera estadísticamente significativa el FSIQ.

4.3. Indicadores de Cuidado familiar (FCI), talla para la edad y desnutrición crónica

El FCI presentó una correlación muy baja con el FSIQ pero estadísticamente significativa en niños menores de 19 meses (r =0,183; P <0,01), que luego pasó a ser baja en el grupo etario de 19 a 30 meses (r =0,362; P <0,001). Esta correlación es similar a las de las escalas de desarrollo cognitivo y de lenguaje del Bayley-III y significativamente superior a aquella de las escalas de desarrollo cognitivo del ASQ-3 y el Denver-II. Cabe mencionar que este aumento fue estadísticamente significativo (Tabla A3). En niños de 31 a 42 meses, la correlación del FCI con el FSIQ continuó siendo baja (r =0,329; P <0,001) y fue similar a la de las otras pruebas cortas. Sin embargo, fue significativamente más baja que la correlación que presentó la escala de desarrollo cognitivo del Bayley-III (P =0,035, Tabla 4).

A pesar de que la validez predictiva de la talla para la edad y la desnutrición crónica aumentó con la edad, este incremento no fue estadísticamente significativo (Tabla A3). Tampoco lo fue

en el grupo etario de 6 a 18 meses. En niños de 19 a 30 meses, la validez predictiva de ambas también fue baja (r =0,164 para la talla para la edad; r =-0,179 para la desnutrición crónica; ambas P <0,001), y estas presentaron una capacidad predictiva significativamente menor que la del FCI (P =0,012 para la desnutrición crónica, P =0,004 para la talla para la edad, Tabla 4). En niños de 31 a 42 meses, la validez predictiva aumentó levemente hasta alcanzar niveles bajos (r =0,203 para la talla para la edad; r = 0,200 para la desnutrición crónica, ambas P <0,001). Para este grupo etario, los resultados de la talla para la edad continuaron siendo significativamente más bajos que los de las escalas de desarrollo cognitivo y de lenguaje del Bayley-III y del Denver-II (P =0,001, Tabla 4), pero no en comparación con los del FCI.

4.4. Análisis en intervalos de seis meses de edad

Los análisis llevados a cabo en intervalos de 6 meses de edad en menores de 19 meses con respecto al Bayley-III, el FCI, la talla para la edad y la desnutrición crónica mostraron que ninguna de estas pruebas fue capaz de predecir significativamente el FSIQ antes de los 12 meses. En niños de entre 13 y 18 meses, se observó un aumento en las correlaciones entre el FSIQ y las escalas de desarrollo cognitivo, lenguaje receptivo y motricidad fina del Bayley-III (r =0,231, P <0,05; r =0,204, P <0,1; r =0,190, P <0,1, respectivamente). El FCI también demostró ser predictivo de desarrollo futuro (r =0,240, P <0,05), pero ni la talla para la edad ni la desnutrición crónica lo fueron.

Se rehicieron todos los análisis utilizando los puntajes del FSIQ estandarizados externamente, los puntajes del FSIQ estandarizados internamente mediante métodos tradicionales, los puntajes del ASQ-3 de los cuestionarios de seis ítems originales, eliminando los valores extremos en la distribución del FSIQ (valores por debajo de -2 DE), y ponderando por la probabilidad inversa de ser reevaluado en la infancia intermedia a fin de ajustar por la pérdida de muestra en el seguimiento. En ninguno de estos análisis de robustez, los resultados se vieron cualitativamente alterados.

5. Discusión

Tal como se esperaba, por lo general, el Bayley-III presentó los valores de validez predictiva más altos. Dicho esto, estos no fueron significativos en menores de 12 meses. En niños de entre 13 y 18 meses, estos valores fueron bajos, al igual que en el grupo etario de 19 a 30 meses; sin embargo, pasaron a ser moderados en el grupo de 31 a 42 meses. Se sabe, por estudios anteriores, que las correlaciones entre las pruebas estandarizadas administradas antes de los 24 meses y las habilidades del niño más tarde en su niñez suelen ser bajas (Bracken 2017; Snow y Hemel 2008). Asimismo, los hallazgos de este estudio son comparables con los reportados en estudios y publicaciones anteriores (Colombo 1993; Fernald et al. 2017; Hamadani et al. 2010, 2013; Pollitt y Triana 1999).

La muestra estaba bien distribuida entre los tres grupos de análisis divididos por rangos de 12 meses de edad. Se observó que, por lo general, la validez predictiva aumenta con la edad, con diferencias particularmente significativas en niños de 6 a 18 meses y de 30 a 42 meses en los casos del Bayley-III y el Denver-II. La validez predictiva de todas las pruebas cortas fue muy baja antes de los 19 meses de edad y, por lo tanto, estas tienen poco valor como predictores del desarrollo futuro. En niños de 19 a 30 meses, las escalas de desarrollo cognitivo y de lenguaje del Denver-II y el BDI-2 mostraron una correlación baja con las del Bayley-III, y las correlaciones entre las escalas de desarrollo cognitivo de las tres pruebas fueron significativamente superiores a la del ASQ-3. En niños de 31 a 42 meses, las

correlaciones del Denver-II y el Bayley-III mostraron valores moderados similares, mientras que en el caso del BDI-2 y el ASQ-3, estos valores fueron significativamente más bajos que los del Bayley-III.

Por otra parte, la capacidad predictiva del WHO-Motor con respecto al CI y el desempeño académico fue prácticamente nula. Contrariamente, en Bangladés, la edad a la que se lograron los hitos del desarrollo motor presentó correlaciones bajas pero significativas con respecto al CI a los 5 años (Hamadani et al. 2013). No obstante, la edad a la que se observaban estos logros fue reportada por las madres durante todo el primer año de vida del niño, lo que podría ofrecer una mayor precisión que llevar a cabo una única evaluación aislada en un momento de tiempo determinado. En los países de ingresos altos, los puntajes del desarrollo motor grueso temprano también mostraron asociaciones más bajas y peor capacidad predictiva del desarrollo futuro que las escalas de desarrollo cognitivo y de lenguaje (Halle et al. 2012).

La validez predictiva del SFII fue similar a la de las escalas de lenguaje y desarrollo cognitivo del Bayley-III, en niños de 19 a 30 meses, edad en la que la adquisición de vocabulario va en aumento. En Bangladés, una prueba de vocabulario similar administrada en niños de 18 meses también mostró una capacidad predictiva comparable con respecto al CI a los 5 años de edad (Hamadani et al. 2010). Cuando la administración de la prueba se basa en los reportes de la madre, no es necesario que el niño interactúe con el encuestador, lo que puede resultar ventajoso, dado que en países de ingresos bajos y medios los niños suelen cohibirse ante extraños. Esto puede ser un elemento a favor del SFII. Sin embargo, y si bien el SFII se encuentra disponible en muchos idiomas (Inventories s. f.), será necesario desarrollar nuevos inventarios en otros idiomas y adaptar los inventarios en idiomas existentes al contexto local para su uso adecuado. En los EE. UU., el SFI y SFII a los 24 meses también fueron predictivos respecto al desarrollo del lenguaje, al igual que el Bayley-III, a pesar de que la validez predictiva de estos variaba según el contexto socioeconómico (Pan et al. 2004).

La elección en torno a qué prueba utilizar debería basarse en los costos, el tiempo de administración y las habilidades requeridas para administrarla, así como también en su validez concurrente y predictiva. Para niños de más de 18 meses, de las pruebas evaluadas, el Denver-II pareció ser la mejor candidata para implementarse a escala, puesto que mostró una capacidad predictiva similar a la del Bayley-III, a pesar de que esta haya sido de baja a moderada, tal como se mencionó anteriormente. La administración del BDI-2, por su parte, llevó demasiado tiempo; y el ASQ-3 demostró una validez baja en niños menores de 31 meses. El tiempo de administración del Denver-II fue de alrededor de 27 minutos, aproximadamente un tercio de lo que demora administrar el Bayley-III. No obstante, este tiempo de administración podría seguir siendo demasiado largo para un estudio a escala. A pesar de que, por lo general, se prefiere utilizar pruebas multidimensionales (Fernandes et al. 2014), en especial cuando los recursos son limitados, una posible solución podría ser utilizar solo las escalas de desarrollo cognitivo y de lenguaje de la prueba seleccionada para disminuir el tiempo de evaluación, dado que estos dominios son, por lo general, los más afectados por las condiciones de pobreza (Rubio-Codina et al. 2015; Rubio-Codina y Grantham-McGregor 2019) y además debido a que estas escalas presentan una mayor validez predictiva. Otra alternativa sería utilizar pruebas de un solo dominio. Por ejemplo, el SFII puede ser útil en niños de 19 a 30 meses si se encuentra disponible en el idioma local.

Sin embargo, se deben realizar estudios adicionales para extender los resultados de este estudio a niños de 36 meses y para desarrollar versiones de estos inventarios en otros idiomas.

La elección de la prueba también dependerá de los objetivos de la encuesta (Rubio-Codina, Araujo, Attanasio, Muñoz, et al. 2016). Los reportes del cuidador podrían ser una mejor opción para evaluar indicadores a nivel poblacional; sin embargo, podrían ser menos convenientes para evaluar programas de estimulación psicosocial y promoción del desarrollo, puesto que podrían presentar sesgo de "observación" (*observation bias*) si, como resultado de la intervención, las madres del grupo de tratamiento pasan más tiempo con el niño y están más conscientes acerca de su proceso de desarrollo o los logros alcanzados con respecto a los diferentes hitos. Asimismo, las madres de los participantes podrían tener un interés sesgado en responder de forma más optimista ante el desarrollo de sus hijos a fin de reportar el éxito de la intervención (sesgo de "deseabilidad" o *desirability bias*). De modo similar, para evaluar intervenciones de nutrición podría ser favorable utilizar una escala de motricidad gruesa, lo cual podría no resultar apropiado para evaluar intervenciones de estimulación psicosocial, debido a la baja capacidad predictiva de estas escalas en relación con el desempeño intelectual futuro, comparadas con las escalas de desarrollo cognitivo y de lenguaje. Si bien tanto el Denver-II como el SFII han demostrado ser sensibles al impacto de programas de transferencias monetarias en Nicaragua (Macours et al. 2012) y Ecuador (Fernald y Hidrobo 2011), respectivamente, sería de gran utilidad llevar a cabo estudios adicionales en torno a la sensibilidad a intervenciones de todas las pruebas cortas.

Es importante destacar que, en niños menores de 31 meses, el FCI presenta una capacidad predictiva similar o superior a la de cualquier otra prueba, incluida el Bayley-III. Además, es una prueba gratuita, rápida (10 a 15 minutos) y fácil de administrar, que proporciona información sobre actividades útiles para los padres (aunque no relacionadas con el cuidado afectivo) y que ha sido ampliamente utilizada en encuestas internacionales —en particular, en las Encuestas de Indicadores Múltiples por Conglomerados (MICS-ECDI s. f.) de UNICEF. A pesar de que el desempeño respecto a un ítem individual puede variar según el contexto o la edad del niño, la calidad del ambiente en el hogar se ha identificado como un factor de protección relevante (Bornstein et al. 2012; Engle et al. 2011) y en muchas ocasiones los resultados en el FCI han mejorado con intervenciones de DIT (Attanasio et al. 2013; Hamadani et al. 2019; Tofail et al. 2013). Creemos que al complementar las escalas de desarrollo cognitivo y de lenguaje del Denver-II con el FCI, tal como se sugirió más arriba, se podría aumentar la sensibilidad en la evaluación de programas y, por lo tanto, sería oportuno investigar esto en mayor detalle en otros estudios.

Asimismo, si bien se ha utilizado la desnutrición crónica como indicador de un desarrollo infantil inadecuado a nivel mundial (Black et al. 2017; Grantham-McGregor et al. 2007), en esta población, el FCI resultó ser un mejor predictor del funcionamiento intelectual general y del desempeño académico futuros. Si estos mismos hallazgos se replican en países con diferentes niveles de la calidad del ambiente en el hogar y con distinto grado de prevalencia y gravedad de la desnutrición crónica, la combinación de ambos podría constituir un indicador a nivel poblacional más efectivo.

Por último, los hallazgos mencionados anteriormente podrían extrapolarse a zonas urbanas de Colombia y posiblemente a zonas urbanas de otros países de América Latina. Sobre la base de nuestro estudio anterior respecto a la validez concurrente de estas pruebas

(Rubio-Codina, Araujo, Attanasio, Muñoz, et al. 2016), se ha demostrado que el Denver-II también resulta apropiado para ser utilizado en Brasil (López Boo, Cubides Mateus, y Llonch Sabatés 2020). Sería necesario llevar a cabo estudios adicionales antes de extrapolar los resultados de este estudio y utilizar estas pruebas en zonas rurales y en otros países de ingresos bajos y medios.

Se limitó la cantidad de pruebas cortas analizadas en este estudio debido a restricciones presupuestarias y de tiempo. Por otro lado, de modo similar, se podrían también evaluar otras pruebas cortas más adecuadas para África y Asia, tales como el instrumento de evaluación del desarrollo *Malawi Developmental Assessment Tool* (MDAT, Gladstone et al. 2010). Además, en la actualidad, se están desarrollando varias pruebas, entre las que se incluyen instrumentos a nivel poblacional e individual dirigidos a niños de 0 a 36 meses (Cavallera et al. 2019; McCoy et al. 2016), a partir de un nuevo enfoque, el puntaje de desarrollo D (*D-Score*), que sintetiza el desarrollo general utilizando una escala de un único intervalo (Weber et al. 2019). Luego de una serie de estudios piloto de validación, estas pruebas podrían resultar apropiadas para ser utilizadas a nivel global (Cavallera et al. 2019).

Un aspecto importante es que algunas de estas pruebas cortas —en especial las de tamizaje— han demostrado una buena sensibilidad y especificidad para identificar niños en alto riesgo de padecer algún retraso en el desarrollo o alguna discapacidad en países de ingresos bajos y medios (Schonhaut et al. 2013). Por cuestiones de diseño, este estudio no aborda este aspecto: la muestra no era lo suficientemente grande, no era representativa de las poblaciones en alto riesgo (p. ej., niños prematuros o con bajo peso al nacer) y no incluyó a niños con alguna discapacidad aparente (Rubio-Codina, Araujo, Attanasio, Muñoz, et al. 2016). En consecuencia, nuestros hallazgos y recomendaciones no pueden generalizarse a estos subgrupos poblacionales.

Entre las limitaciones de este estudio se incluyen la alta pérdida de muestra entre rondas y la falta de estandarización para Colombia de las pruebas usadas en la infancia intermedia. Sin embargo, ponderar por la pérdida de muestra en los análisis no alteró los resultados. Asimismo, las pruebas para medir el desarrollo en la infancia intermedia mostraron buenos niveles de confiabilidad y presentaron correlaciones adecuadas entre sí, al igual que con las características socioeconómicas de los niños y sus hogares y con medidas del desarrollo en la primera infancia, por lo que demostraron ser válidas para esta población. Otro problema es que el Denver-II no incluye por separado una escala de desarrollo cognitivo per se. Sin embargo, la escala de motricidad fina adaptativa combina ítems relacionados con la cognición y la motricidad fina, y hemos demostrado que esta escala presenta una buena correlación con el CI futuro, mejor o similar a las de las escalas de desarrollo cognitivo de otras pruebas cortas y semejante a la del Bayley-III en niños de 19 meses en adelante. Otra limitación es que la validez predictiva de las pruebas de desarrollo temprano podría confundirse, en parte, con el desempeño de las pruebas que miden el desarrollo en la infancia intermedia, puesto que todos los niños se evaluaron 5,5 años después de la evaluación inicial. El desempeño de estas pruebas parece aumentar con la edad de acuerdo con ciertos indicadores de confiabilidad y validez (correlaciones de las pruebas entre sí y con variables socioeconómicas), pero no así en relación con otros indicadores (consistencia interna).

El hecho de que la muestra de niños evaluados en la primera infancia y en la infancia intermedia sea de gran tamaño y con representatividad de la población de estudio —aunque esta población sea urbana— son dos de las importantes ventajas que presenta este estudio,

al igual que el número de pruebas analizadas y la calidad de la prueba de referencia en la primera infancia (Bayley-III) y de la prueba para la medición del CI en la infancia intermedia (WISC-V).

6. Conclusiones

En general, la validez predictiva de todas las pruebas aumentó con la edad. Las escalas de lenguaje y de desarrollo cognitivo presentaron los valores de validez predictiva más elevados. Asimismo, ninguna de las pruebas demostró una capacidad predictiva significativa en niños menores de 19 meses y, a partir de esta edad, la validez predictiva de las pruebas solo fue de baja a moderada en el mejor de los casos. De las pruebas cortas analizadas, el SFII en niños de 19 a 30 meses y el Denver-II en niños de 19 a 42 meses demostraron ser las pruebas más factibles de usar y con mayor validez. En niños menores de 31 meses, el FCI fue un buen predictor al igual que el Bayley-III, y mostró mejores resultados en comparación con el indicador de desnutrición crónica. Estos hallazgos sugieren que combinar el FCI con las escalas de lenguaje y de desarrollo cognitivo del Denver-II podría ser una buena alternativa para evaluar las intervenciones a escala destinadas a promover el DIT en niños menores de 36 meses. Asimismo, añadir el FCI al indicador de desnutrición crónica —como indicador poblacional *aproximado* (*proxy*) de los niveles de desarrollo infantil— también podría ser una forma de lograr estimar mejor la cantidad de niños que se encuentran en riego de padecer rezagos en su desarrollo a nivel global. Es importante que estos hallazgos se repliquen en otras regiones y que estos estudios futuros incluyan otras pruebas recientemente desarrolladas o en curso de desarrollo.

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, Orazio, Camila Fernández, Emla Fitzsimons, Sally Grantham-McGregor, Costas Meghir, y Marta Rubio-Codina. 2013. “Enriching the home environment of low-income families in Colombia: a strategy to promote child development at scale.” *Early Childhood Maters* 35–39.

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

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

Black, Maureen M., Susan P. Walker, Lia C. Fernald, Christopher T. Andersen, Ann M. DiGirolamo, Chunling Lu, Dana C. McCoy, Fink Gunther, Yusra R. Shawar, Jeremy Shiffman, Amanda E. Devercelli, Quentin T. Wodon, Emily Vargas-Baron, Sally Grantham-McGregor, y Lancet Early Childhood Development Series Steering Committee. 2017. “Early childhood development coming of age: science through the life course”. *The Lancet* 389(10064):77–90.

Bornstein, Marc H., Pia Rebello Britto, Yuko Nonoyama-Tarumi, Yumiko Ota, Oliver Petrovic, y Diane L. Putnick. 2012. “Child development in developing countries: Introduction and methods”. *Child Development* 83(1):16–31.

Bracken, Bruce A. 2017. *Creating the Optimal Preschool Testing Situation. Psychoeducational Assess- ment of Preschool Children*. editado por B. A. B. and R. Nagle. Mahwah, NJ: Lawrence Erlbaum Associates.

Bradley, Robert H., Bettye M. Caldwell, Stephen L. Rock, Holly M. Hamrick, y Pandia Harris. 1988. “Home Observation for Measurement of the Environment: Development of a Home Inventory for use with families having children 6 to 10 years old”. *Contemporary Educational Psychology* 13(1):58–71.

Caldwell, Bettye M., y Robert H. Bradley. 2003. *HOME Inventory Administration Manual. Comprehensive Edition.* editado por U. of Arkansas. Print Design, Little Rock.

Cavallera, Vanessa, Maureen Black, Kieran Bromley, Jorge Cuartas, Iris Eekhout, Günther Fink, Melissa Gladstone, Katelyn Hepworth, Magdalena Janus, Patricia Kariger, Gillian Lancaster, Dana McCoy, Gareth McCray, Raikes Abbie, Marta Rubio-Codina, Stef van Buuren, Marcus Waldman, Susan Walker, Ann Weber, y Tarun Dua. 2019. “The Global Scale for Early Development (GSED)”. *Early Childhood Matters* 80–84.

Colombo, John. 1993. *Infant cognition: Predicting later intellectual functioning. Individual differences and developmental series.* Newbury Park, CA: Sage Publications, Inc.

Dunn, Lloyd M., y Leota M. Dunn. 1981. *Peabody Picture Vocabulary Test - Revised*. Circle Pines, Minnesota: American Guidance Service.

Dunn, Lloyd M., Eligio R. Padilla, Delia E. Lugo, y Leota M. Dunn. 1986. *Test de Vocabulario en Imágenes Peabody. Adaptación Hispanoamericana*. Minneapolis, MN: NCS Pearson, Inc.

Efron, Bradley. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Vol. 38. Philadelphia, PA: SIAM.

ELCA 2010 & 2013. “https://encuestalongitudinal.uniandes.edu.co/es/”.

Engle, Patrice L., Lia C. H. Fernald, Harold Alderman, Jere Behrman, Chloe O’Gara, Aisha Yousafzai, Meena Cabral De Mello, Melissa Hidrobo, Nurper Ulkuer, Ilgi Ertem, y Selim Iltus. 2011. “Strategies for reducing inequalities and improving developmental outcomes for young children in low-income and middle-income countries”. *The Lancet* 378(9799):1339–53.

Evans, J. D. 1996. *Straightforward statistics for the behavioral sciences*. editado por B. Publishing. Pacific Grove, CA.

Fenson, L., PS Dale, JS Reznick, D. Thal, E. Bates, JP Hartung, SJ Pethick, y JS Reilly. 2002. *The MacArthur Communicative Development Inventories: users guide and technical manual*. Baltimore, MD: Paul Brookes Publishing Co.

Fernald, L. C. H., 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, Lia C. H., Elizabeth Prado, Patricia Kariger, y Abbie Raikes. 2017. *A Toolkit for Measuring Early Childhood Development in Low and Middle-Income Countries*. 122031. Washington, DC.

Fernald, Lia C., y Melissa 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.

Fernandes, Michelle, Alan Stein, Charles R. Newton, Leila Cheikh-Ismail, Michael Kihara, Katharina Wulff, Enrique de León Quintana, Luis Aranzeta, Aureli Soria-Frisch, Javier Acedo, David Ibanez, Amina Abubakar, Francesca Giuliani, Tamsin Lewis, Stephen Kennedy, y Jose Villar. 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, William K., Josiah Dodds, Philip Archer, Beverly Bresnick, Patrick Maschka, Norma Edelmann, y Howard Shapiro. 1990. *The DENVER II Technical Manual*. Denver, CO: Denver Developmental Materials.

Frankenburg, William K., Josiah Dodds, Philip Archer, Howard Shapiro, y Beverly Bresnick. 1992. “A major revision and restandardization of the Denver Developmental Screening Test”. *Pediatrics* 89:91–97.

Gladstone, Melissa, Gillian A. Lancaster, Eric Umar, Maggie Nyirenda, Edith Kayira, Nynke R. Van, Den Broek, y Rosalind L. Smyth. 2010. “The Malawi Developmental Assessment Tool (MDAT): The Creation, Validation, and Reliability of a Tool to Assess Child Development in Rural African Settings”. 7(5).

Grantham-McGregor, Sally, Yin Bun Cheung, Santiago Cueto, Paul Glewwe, Linda Richter, y Barbara Strupp. 2007. “Developmental potential in the first 5 years for children in developing countries”. *The Lancet* 369(9555):60–70.

Halle, Tamara G., Elizabeth C. Hair, Margaret Burchinal, Rachel Anderson, y Martha Zaslow. 2012. *In the Running for Successful Outcomes: Exploring the Evidence for Thresholds of School Readiness.* Washington, DC.

Hamadani, Jena D., Helen Baker-Henningham, Fahmida Tofail, Fardina Mehrin, Syed N.

Huda, y Sally 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, Jena D., Syed N. Huda, Fahmida Khatun, y Sally M. Grantham-McGregor. 2006. “Psychosocial stimulation improves the development of undernourished children in rural Bangladesh.” *The Journal of nutrition* 136(10):2645–52.

Hamadani, Jena D., Syeda F. Mehrin, Fahmida Tofail, Mohammad I. Hasan, Syed N. Huda, Helen Baker-Henningham, Deborah Ridout, y Sally Grantham-McGregor. 2019. “Integrating an early childhood development programme into Bangladeshi primary health-care services: an open-label, cluster-randomised controlled trial”. *The Lancet Global Health* 7(3):e366–75.

Hamadani, Jena Derakhshani, Fahmida Tofail, Tim Cole, y Sally 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.

Inventories, MacArthur-Bates Communicative Development. s. f. “http://mb-cdi.stanford.edu/adaptations_ol.htm”.

Jackson-Maldonado, Donna, Virginia A. Marchman, y Lia C. H. Fernald. 2012. “Short-form versions of the Spanish MacArthur–Bates Communicative Development Inventories”. *Applied Psycholinguistics* 34(14):837–68.

Jackson-Maldonado, Donna, D. Thal, Virginia 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.

Kariger, Patricia, Edward A. Frongillo, Patrice Engle, Pia M. Rebell. Britto, Sara M. Sywulka, y Purnima Menon. 2012. “Indicators of Family Care for Development for Use in Multicountry Surveys”. *Journal of Health, Population and Nutrition* 30(4):472–86.

López Boo, Florencia, Mayaris Cubides Mateus, y Ana Llonch Sabatés. 2020. “Initial psychometric properties of the Denver II in a sample from Northeast Brazil”. *Infant Behavior and Development* 58(January 2020):101391.

Macours, Karen, Norbert Schady, y Renos 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.

McCoy, Dana Charles, Maureen M. Black, Bernadette Daelmans, y Tarun Dua. 2016. “Measuring development in children from birth to age 3 at population level”. *Early Childhood Matters* 34–39.

MICS-ECDI. s. f. “MICS UNICEF”. Recuperado (http://mics.unicef.org/).

Muñoz-Sandoval, A. F., R. W. Woodcock, K. S. McGrew, y N. Mather. 2005. *Batería III Woodcock-Muñoz: Pruebas de aprovechamiento.* Rolling Meadows, IL: Riverside Publishing.

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):725–31.

Newborg, J. 2005. *Battelle Developmental Inventory-2nd Edition*. Rolling Meadows, IL: Riverside Publishing.

Pan, Barbara Alexander, Meredith L. Rowe, Elizabeth Spier, y Catherine Tamis-Lemonda. 2004. “Measuring productive vocabulary of toddlers in low-income families: Concurrent and predictive validity of three sources of data”. *Journal of Child Language* 31(3):587–608.

Pollitt, Ernesto, y Nina Triana. 1999. “Stability, predictive validity, and sensitivity of mental and motor development scales and pre-school cognitive tests among low-income children in developing countries”. *Food and Nutrition Bulletin* 20(1):45–52.

Rubio-Codina, Marta, M. Caridad Araujo, Orazio Attanasio, y Sally Grantham-McGregor. 2016. *Concurrent Validity and Feasibility of Short Tests Currently Used to Measure Early Childhood Development in Large Scale Studies: Methodology and Results*. IDB-WP-723. Washington, D.C.

Rubio-Codina, Marta, M. Caridad Araujo, Orazio Attanasio, Pablo Muñoz, y Sally Grantham-McGregor. 2016. “Concurrent Validity and Feasibility of Short Tests Currently Used to Measure Early Childhood Development in Large Scale Studies”. *Plos One* 11(8):e0160962.

Rubio-Codina, Marta, Orazio Attanasio, Costas Meghir, Natalia Varela, y Sally Grantham-McGregor. 2015. “The Socioeconomic Gradient of Child Development: Cross-Sectional Evidence from Children 6–42 Months in Bogota”. *Journal of Human Resources* 50(2):464–83.

Rubio-Codina, Marta, y Sally Grantham-McGregor. 2019. “Evolution of the wealth gap in child development and mediating pathways: Evidence from a longitudinal study in Bogota, Colombia”. *Developmental Science* (e12810.):1–15.

Schonhaut, L., I. Armijo, M. Schonstedt, J. Alvarez, y M. Cordero. 2013. “Validity of the Ages and Stages Questionnaires in Term and Preterm Infants”. *Pediatrics* 131(5):e1468–74.

Snow, Catherine E., y Susan B. Van Hemel. 2008. *Early Childhood Assessment: Why, What, and How*. Washington, DC: The National Academies Press.

Squires, Jane, D. Bricker, E. Twombly, R. Nickel, Jantina Clifford, Kimberly 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.

Tofail, Fahmida, Jena D. Hamadani, Fardina Mehrin, Deborah A. Ridout, Syed N. Huda, y Sally M. Grantham-McGregor. 2013. “Psychosocial Stimulation Benefits Development in Nonanemic Children but Not in Anemic, Iron-deficient Children”. *The Journal of Nutrition* 143(6):885–93.

Weber, Ann M., Marta Rubio-Codina, Susan P. Walker, Stef van Buuren, Iris Eekhout, Sally M. Grantham-McGregor, M. Caridad Araujo, Susan M. Chang, Lia C. H. Fernald, Jena D. Hamadani, Charlotte Hanlon, Simone M. Karam, Betsy Lozoff, Lisy Ratsifandrihamanana, Linda M. Richter, Maureen M. Black, y Global Child Development Collaborators. 2019. “The D-score: a metric for interpreting the early development of infants and toddlers across global settings”. *BMJ Global Health* 4:e001724.

Wechsler, David. 2014. *WISC-V. Wechsler Intelligence Scale for Children —Fifth Edition*. Bloomington, MN: NCS Pearson, Inc.

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.

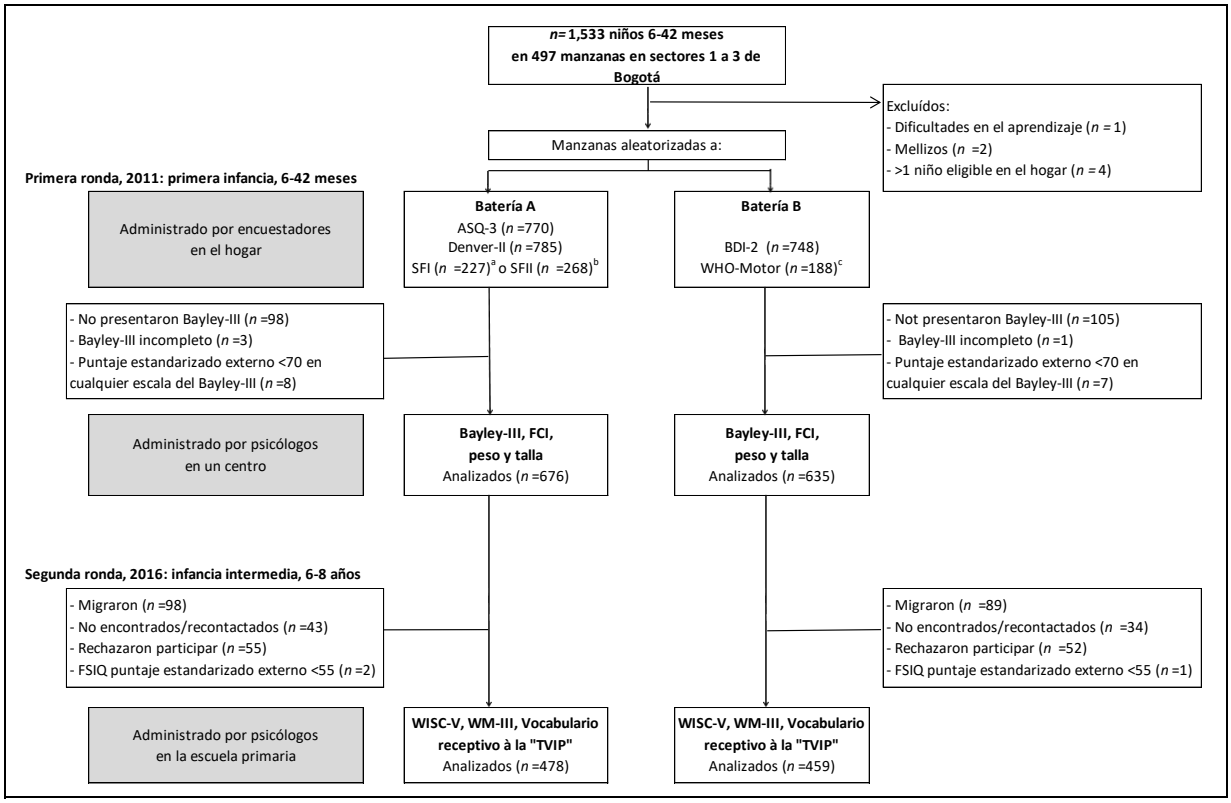
Wijnhoven, Trudy MA, Mercedes de Onis, Adelheid W. Onyango, Tracey Wang, Gunn-Ellin A. Bjoerneboe, Nita Bhandari, Anna Lartey, y Badriya Al Rashidi. 2004. “Asessment of gross motor development in the WHO Multicentre Growth Reference Study”. *Food and Nutrition Bulletin* 25(1):S37–45.

Woodcock, R. W., K. S. McGrew, y N. Mather. 2001. *Woodcock-Johnson III.* Itasca, IL: Riverside Publishing.

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

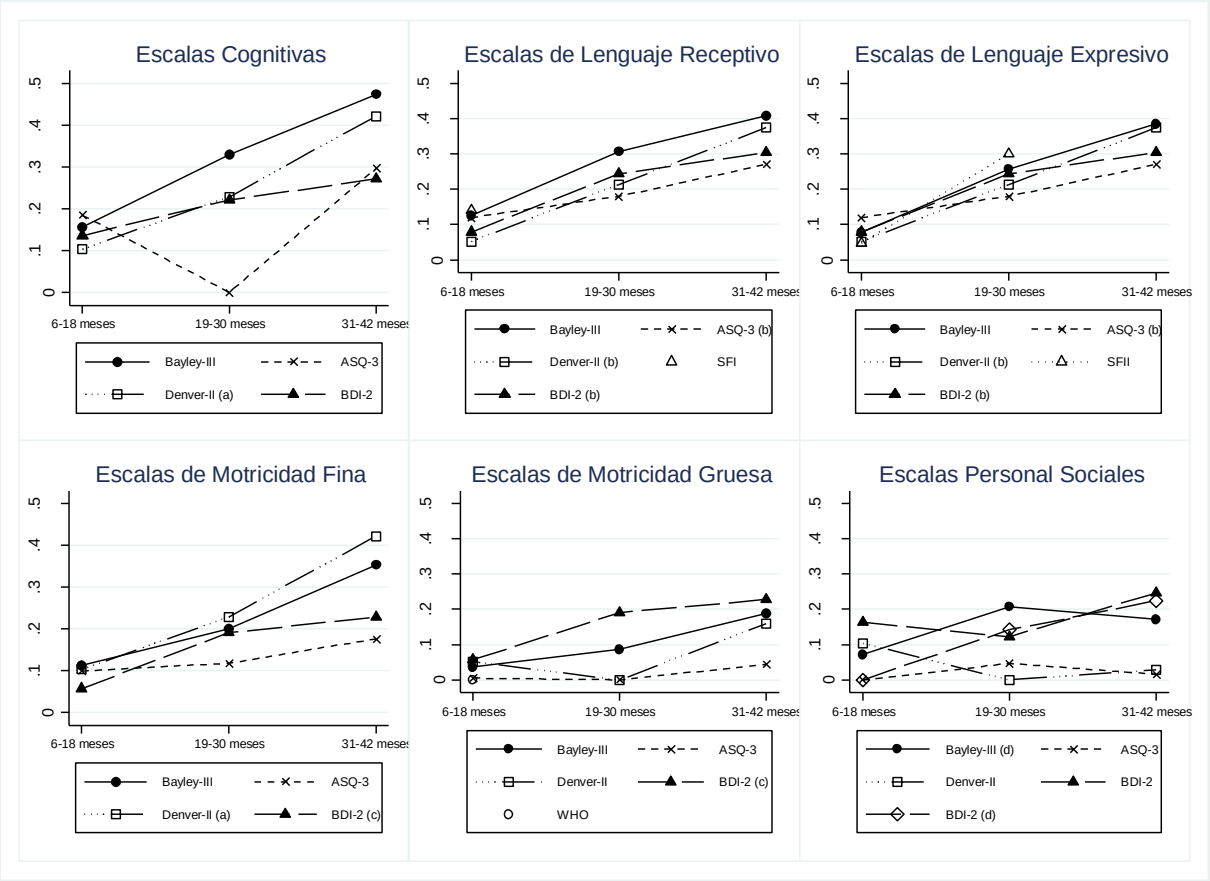
Gráficos y tablas

Gráfico 1. Diseño del estudio y diagrama del flujo de los participantes



Notas: ASQ-3 (Cuestionarios de Edades y Etapas, tercera edición); Denver-II (Prueba de Tamizaje del Desarrollo de Denver, segunda edición); SFI (Inventarios MacArthur-Bates del Desarrollo de Habilidades Comunicativas, versión abreviada I en español); SFII (Inventarios MacArthur-Bates del Desarrollo de Habilidades Comunicativas, versión abreviada II en español); BDI-2 (prueba de tamizaje Inventario del Desarrollo de Battelle, segunda edición); WHO-Motor (Hitos del Desarrollo Motor Grueso de la Organización Mundial de la Salud); Bayley-III (Escalas Bayley de Desarrollo Infantil, tercera edición); FCI (Indicadores de Cuidado Familiar); WISC-V (Escala Wechsler de Inteligencia para Niños, quinta edición); WM-III (Prueba de Aprovechamiento de Woodcock-Muñoz, tercera edición); TVIP (Test de Vocabulario en Imágenes de Peabody, revisado, versión en español, conjunto de palabras seleccionadas y adaptadas). [a] Niños de 8 a 18 meses. [b] Niños de 19 a 30 meses. [c] Niños de 6 a 15 meses.

Gráfico 2. Validez predictiva del Bayley-III y de las pruebas cortas entre los 6 y 42 meses del Coeficiente Intelectual (FSIQ) entre los 6 y 8 años, por grupo etario y dominio



Notas: Cada cuadrante del Gráfico 2 muestra, para cada dominio del desarrollo, la correlación promedio entre los puntajes obtenidos en la primera infancia y aquellos correspondientes al FSIQ en la infancia intermedia, por grupo etario. El grupo etario más joven en la primera infancia es el de 6 a 18 meses y en la infancia intermedia, el de 6 años; el grupo etario intermedio en la primera infancia es el de 19 a 30 meses y en la infancia intermedia, el de 7 años; y el grupo etario de mayor edad en la primera infancia es el de 31 a 42 meses y en la infancia intermedia, el de 8 años. Los tamaños de muestra para cada correlación (punto ilustrado) se presentan en la Tabla 2. Véase el Gráfico 1 para obtener las referencias de los acrónimos utilizados. [a] El Denver-II corresponde a la escala de motricidad fina adaptativa; [b] el ASQ-3, el Denver-II y el BDI-2 corresponden a las escalas de comunicación/lenguaje; [c] el BDI-2 de desarrollo motor combina ítems de motricidad fina y gruesa; [d] el Bayley-III corresponde a la escala socioemocional; y el BDI-2 corresponde a la escala de habilidades adaptativas.

Tabla 1. Características de los niños evaluados en la infancia intermedia y de sus familias, por batería de pruebas administrada en la primera infancia

	Batería A (n_A=478)	Batería B (n_B =459)	P-Valor de la diferencia entre las baterías
I. Características del niño			
Edad del niño al ingreso en el estudio, %			
6-18 meses	31,6	33,1	0,64
19-30 meses	36,2	35,5	0,822
31-42 meses	32,2	31,4	0,787
Edad del niño en la infancia intermedia, %			
6 años	33,1	33,1	0,985
7 años	35,4	36,6	0,698
8 años	31,6	30,3	0,674
Niñas, %	45,4	49,7	0,173
Prematuro (edad gestacional <37 semanas), %	14,6	15,5	0,754
Talla para la edad[a] al ingreso, puntaje z, media (DE)	-1,1 (1,1)	-1,1 (1,1)	0,319
Desnutrición crónica[a] (talla para la edad <-2 DE) al ingreso, %	16,6	17,9	0,595
II. Características de los padres en la infancia intermedia			
Edad de la madre[a], media (DE)	33,3 (6,9)	32,3 (6,5)	0,037
Educación de la madre, media (DE)	10,9 (3,2)	11,2 (3,3)	0,241
Educación del padre[a], media (DE)	8,9 (4,1)	9,3 (4,0)	0,087
III. Características del hogar y del ambiente			
Estrato socioeconómico en la infancia intermedia, %			
1. (25% más pobre)	27,2	27,5	0,93
2	39,7	42,5	0,422
3	32	28,5	0,329
4. (25% más rico)	1	1,5	0,598
Tamaño del hogar en la infancia intermedia, media (DE)	4,5 (1,4)	4,3 (1,5)	0,133
Índice de riqueza del hogar en la infancia intermedia, media (DE)	-0,01 (1,03)	0,03 (0,95)	0,52
Variedad en los materiales de juego (FCI) en la primera infancia, media (DE)	5,0 (2,3)	5,0 (2,3)	0,143
Variedad en las actividades de juego (FCI) en la primera infancia, media (DE)	3,9 (1,9)	3,7 (1,8)	0,955
Puntaje total del FCI (materiales y actividades de juego) en la primera infancia, estandarizado internamente, media (DE)	0,16 (1,7)	0,13 (1,7)	0,786
Puntaje total del MC-HOME en la infancia intermedia, estandarizado internamente, media (DE)	0,05 (0,9)	-0,05 (1,0)	0,145
IV. Desarrollo del niño en la primera infancia			
Bayley-III, puntajes estandarizados internamente, media (DE)			
Desarrollo cognitivo	0,03 (0,10)	0,05 (0,99)	0,763
Lenguaje receptivo	0,08 (0,99)	0,02 (1,02)	0,346
Lenguaje expresivo	0,04 (0,99)	0,02 (1,00)	0,713
Motricidad fina	0,05 (0,98)	0,00 (0,99)	0,39
Motricidad gruesa	0,03 (1,02)	0,01 (0,10)	0,758
Socioemocional	0,02 (0,98)	0,00 (0,97)	0,702
V. Desarrollo del niño en la infancia intermedia			
FSIQ, WISC-V, estandarizado externamente, media (DE)	88,6 (12,1)	88,6 (12,6)	0,991
FSIQ, WISC-V, estandarizado internamente, media (DE)	0,12 (4,40)	-0,04 (4,50)	0,605
Puntaje de desempeño académico, estandarizado internamente, media (DE)	0,03 (2,39)	0,01 (2,39)	0,903

Notas: [a] Datos incompletos para algunas de las variables. Los tamaños de la muestra para estos son: talla para la edad y desnutrición crónica (n_A=477); edad de la madre (n_A=453, n_B=433); educación del padre (n_A=456, n_B=424). DE significa desviación estándar.

Tabla 2. Correlaciones de las pruebas en la infancia intermedia entre sí y con variables socioeconómicas concurrentes

	FSIQ WISC-V	Puntaje de desempeño académico	Aritmética, WM-III	Comprensión lectora, WM-III	Vocabulário receptivo, à la "TVIP"
Variables socioeconómicas en la infancia intermedia					
Educación de la madre	0,315***	0,288***	0,172***	0,292***	0,226***
Índice de riqueza del hogar	0,308***	0,300***	0,212***	0,285***	0,221***
Puntaje total del MC-HOME, estandarizado internamente	0,331***	0,342***	0,266***	0,289***	0,264***
Pruebas en la infancia intermedia, estandarizadas internamente					
FSIQ, WISC-V	1				
Puntaje de desempeño académico	0,706***	1			
Aritmética, WM-III	0,531***	0,798***	1		
Comprensión lectora, WM-III	0,600***	0,824***	0,512***	1	
Vocabulario receptivo, basado en el TVIP	0,563***	0,776***	0,402***	0,461***	1

Notas: N = 937 niños. Correlaciones de Pearson entre los puntajes de las pruebas estandarizados internamente (netos de los efectos del evaluador/encuestador). Los errores estándar (EE) se calcularon mediante métodos *Bootstrap*, estratificando por los estratos del diseño: grupo etario y sector socioeconómico (n =2000 repeticiones).

Tabla 3. Validez predictiva del Bayley-III, las pruebas cortas, el FCI, la talla para la edad y la desnutrición crónica entre los 6 y 42 meses del Coeficiente Intelectual (FSIQ) y el desempeño académico entre los 6 y 8 años, por edad en la primera infancia

	6-18 meses en la primera infancia		19-30 meses en la primera infancia		31-42 meses en la primera infancia	
Pruebas en la primera infancia, 6-42 meses	**FSIQ (CI)**	**Desempeño Académico**	**FSIQ (CI)**	**Desempeño Académico**	**FSIQ (CI)**	**Desempeño Académico**
Bayley-III	*n = 303*		*n = 336*		*n = 298*	
Desarrollo cognitivo	0,157**	0,135*	0,330***	0,309***	0,474***	0,436***
Lenguaje receptivo	0,126*	0,150**	0,307***	0,348***	0,409***	0,397***
Lenguaje expresivo	0,079	0,089	0,257***	0,308***	0,386***	0,398***
Motricidad fina	0,113*	0,122*	0,201***	0,213***	0,353***	0,311***
Motricidad gruesa	0,036	-0,071	0,087	0,143**	0,188**	0,175**
Socioemocional	0,073	0,019	0,207***	0,208***	0,172**	0,119*
ASQ-3 (adaptado)	*n = 145*		*n = 172*		*n = 153*	
Resolución de problemas	0,185**	0,027	-0,122	-0,008	0,297***	0,310***
Comunicación	0,120	0,112	0,180*	0,223**	0,271***	0,317***
Motricidad fina	0,099	0,038	0,118	0,167*	0,176**	0,247***
Motricidad gruesa	0,005	-0,065	-0,099	-0,110	0,044	0,038
Personal-social	-0,115	-0,021	0,047	0,057	0,016	0,009
Denver-II	*n = 148*		*n = 169*		*n = 148*	
Lenguaje	0,052	-0,026	0,214**	0,286***	0,375***	0,373***
Motricidad fina adaptativa	0,103	0,116	0,229**	0,146*	0,422***	0,438***
Motricidad gruesa	0,052	-0,082	-0,005	0,035	0,160*	0,190*
Personal-social	0,104	0,015	-0,031	0,049	0,028	0,058
BDI-2 (Battelle)	*n = 152*		*n = 163*		*n = 144*	
Desarrollo cognitivo	0,136	0,175*	0,221**	0,238**	0,272*	0,305***
Comunicación	0,079	0,095	0,244**	0,326***	0,305***	0,318***
Motricidad	0,057	-0,010	0,191*	0,136	0,228*	0,220*
Personal-social	0,163*	0,125	0,122	0,181**	0,247**	0,224**
Habilidades adaptativas	-0,058	-0,033	0,142	0,083	0,224**	0,192*
SFI y SFII (MacArthur)	*n = 126*[a]		*n = 172*			
Lenguaje receptivo	0,140	0,181*				
Lenguaje expresivo	0,047	0,055	0,301***	0,298***		
WHO-Motor	*n = 110*[b]					
Motricidad gruesa	-0,033	-0,068				
FCI, talla para la edad, desnutrición crónica	*n = 303*		*n = 336*		*n = 298*	
FCI	0,183**	0,180**	0,362***	0,384***	0,329***	0,310***
Talla para la edad	0,085	0,088	0,164**	0,172**	0,203***	0,206***
Desnutrición crónica	-0,054	-0,113*	-0,179***	-0,207***	-0,200***	-0,247***

Notas: Correlaciones de Pearson entre los puntajes estandarizados internamente (netos de los efectos del evaluador/encuestador), a excepción de la talla para la edad y la desnutrición crónica. Los *P*-valores se calcularon mediante métodos *Bootstrap*, estratificando por los estratos del diseño: grupo etario y sector socioeconómico (n =2000 repeticiones). El FCI incluye los materiales y actividades de juego. La desnutrición crónica se define en función de una talla para la edad de -2 desviaciones estándar (DE) por debajo de la mediana establecida por la OMS. * p<0,05, ** p<0,01, *** p<0,001. [a] Niños de 8 a 18 meses; [b] Niños de 6 a 15 meses.

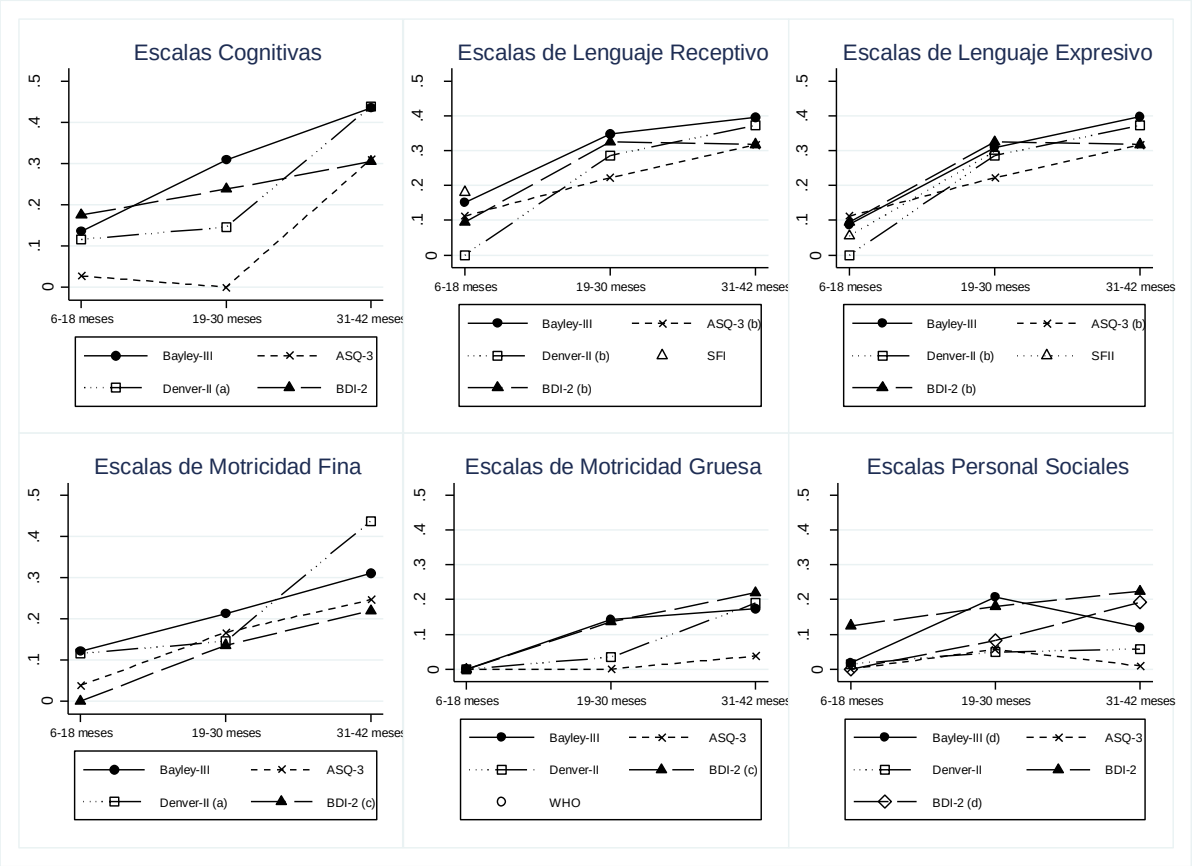
Tabla 4. Comparación de todas las correlaciones significativamente diferentes (*P* <0,05) de las escalas de desarrollo cognitivo y de lenguaje de todas las pruebas cortas, el Bayley-III, el FCI, la talla para la edad y la desnutrición crónica en la primera infancia con el posterior Coeficiente Intelectual (FSIQ) y desempeño académico, por grupo etario

	Coeficiente Intelectual, FSIQ	Desempeño Académico
19-30 meses	Bayley-III, desarrollo cognitivo > ASQ-3, resolución de problemas (*P* <0,0001)	Bayley-III, desarrollo cognitivo > ASQ-3, resolución de problemas (*P* =0,001)
	Bayley-III, desarrollo cognitivo > Talla para la edad (*P* =0,037)	Bayley-III, desarrollo cognitivo > Denver-II, motricidad fina adaptativa (*P* =0,033)
	Denver-II, motricidad fina adaptativa > ASQ-3, resolución de problemas (*P* <0,0001)	Bayley-III, lenguaje expresivo > Talla para la edad (*P* =0,037)
	BDI-2, desarrollo cognitivo > ASQ-3, resolución de problemas (*P* =0,002)	BDI-2, desarrollo cognitivo > ASQ-3, resolución de problemas (*P* =0,030)
	SFII, lenguaje expresivo > ASQ-3, resolución de problemas (*P* <0,0001)	SFII, lenguaje expresivo > ASQ-3, resolución de problemas (*P* =0,003)
	FCI > ASQ-3, resolución de problemas (*P* <0,0001)	FCI > ASQ-3, resolución de problemas (*P* <0,0001)
	FCI > ASQ-3, comunicación (*P* =0,012)	FCI > ASQ-3, comunicación (*P* =0,035)
	FCI > Denver-II, motricidad fina adaptativa (*P* =0,049)	FCI > Denver-II, motricidad fina adaptativa (*P* =0,002)
	FCI > Desnutrición crónica (*P* =0,012)	FCI > Desnutrición crónica (*P* =0,014)
	FCI > Talla para la edad (*P* =0,003)	FCI > Talla para la edad (*P* =0,002)
	Desnutrición crónica > ASQ-3, resolución de problemas (*P* =0,003)	
	Talla para la edad > ASQ-3, resolución de problemas (*P* =0,005)	
31-42 meses	Bayley-III, desarrollo cognitivo > ASQ-3, resolución de problemas (*P* =0,038)	Bayley-III, desarrollo cognitivo > FCI (*P* =0,027)
	Bayley-III, desarrollo cognitivo > ASQ-3, comunicación (*P* =0,033)	Bayley-III, desarrollo cognitivo > Desnutrición crónica (*P* =0,007)
	Bayley-III, desarrollo cognitivo > BDI-2, comunicación (*P* =0,016)	Bayley-III, lenguaje expresivo > Desnutrición crónica (*P* =0,022)
	Bayley-III, desarrollo cognitivo > FCI (*P* =0,035)	Bayley-III, desarrollo cognitivo > Talla para la edad (*P* =0,001)
	Bayley-III, desarrollo cognitivo > Desnutrición crónica (*P* <0,0001)	Bayley-III, lenguaje receptivo > Talla para la edad (*P* =0,017)
	Bayley-III, lenguaje receptivo > Desnutrición crónica (*P* =0,010)	Bayley-III, lenguaje expresivo > Talla para la edad (*P* =0,003)
	Bayley-III, lenguaje expresivo > Desnutrición crónica (*P* =0,007)	Denver-II, motricidad fina adaptativa > Desnutrición crónica (*P* =0,012)
	Bayley-III, desarrollo cognitivo > Talla para la edad (*P* <0,0001)	Denver-II, motricidad fina adaptativa > Talla para la edad (*P* =0,004)
	Bayley-III, lenguaje receptivo > Talla para la edad (*P* =0,010)	
	Bayley-III, lenguaje expresivo > Talla para la edad (*P* =0,008)	
	Denver-II, motricidad fina adaptativa > Desnutrición crónica (*P* =0,004)	
	Denver-II, lenguaje > Desnutrición crónica (*P* =0,046)	
	Denver-II, motricidad fina adaptativa > Talla para la edad (*P* =0,007)	
	Denver-II, lenguaje > Talla para la edad (*P* =0,045)	

Notas: Número de observaciones para cada correlación comparada, tal como en la Tabla 3. Los P-valores se calcularon mediante métodos *Bootstrap*, estratificando por los estratos del diseño: grupo etario y sector socioeconómico (n =2000 repeticiones).

Apéndice: Gráficos y Tablas

Gráfico A1. Validez predictiva del Bayley-III y de las pruebas cortas entre los 6 y 42 meses del desempeño académico entre los 6 y 8 años, por grupo etario y dominio



Notas: Cada cuadrante del Gráfico A1 muestra, para cada dominio del desarrollo, la correlación promedio entre los puntajes obtenidos en la primera infancia y el desempeño académico en la infancia intermedia, por grupo etario. El grupo etario más joven en la primera infancia es el de 6 a 18 meses y en la infancia intermedia, el de 6 años; el grupo etario intermedio en la primera infancia es el de 19 a 30 meses y en la infancia intermedia, el de 7 años; y el grupo etario de mayor edad en la primera infancia es el de 31 a 42 meses y en la infancia intermedia, el de 8 años. Los tamaños de muestra para cada correlación (punto ilustrado) se presentan en la Tabla 2. Véase el Gráfico 1 para obtener las referencias de los acrónimos utilizados. [a]El Denver-II corresponde a la escala de motricidad fina adaptativa; [b]el ASQ-3, el Denver-II y el BDI-2 corresponden a las escalas de comunicación/lenguaje; [c]el BDI-2 de desarrollo motor combina ítems de motricidad fina y gruesa; [d]el Bayley-III corresponde a la escala socioemocional; y el BDI-2 corresponde a la escala de habilidades adaptativas.

Tabla A1. Puntajes crudos para el Bayley-III, las pruebas cortas, el WISC-V y las mediciones del desempeño académico, y puntajes compuestos para el Bayley-III, para la muestra completa y por baterías administradas en la primera infancia

	Todos los niños (N = 937)				Batería A (n_A =478)		Batería B (n_B =459)		*P-valor diferencia entre las baterías*
	Media	**DE**	**mín.**	**máx.**	**Media**	**DE**	**Media**	**DE**	
Bayley-III, puntajes crudos									
Desarrollo cognitivo	59,14	13,84	26	83	59,03	14,34	59,25	13,32	0,809
Lenguaje receptivo	25,75	9,08	9	44	25,63	9,2	25,86	8,96	0,697
Lenguaje expresivo	25,55	10,25	5	47	25,27	10,28	25,85	10,23	0,393
Motricidad fina	39,51	9,93	18	63	39,74	10	39,26	9,85	0,466
Motricidad gruesa	52,66	11,35	20	70	52,42	11,59	52,92	11,09	0,505
Socioemocional	106	30,05	47	170	105,5	30,38	106,4	29,73	0,646
Bayley-III, puntajes compuestos (estandarizados externamente)									
Desarrollo cognitivo	98,61	8,83	70	135	97,94	7,76	99,31	9,78	0,048
Lenguaje	96,94	9,97	71	138	95,87	8,59	98,05	11,13	0,005
Motricidad	99,79	10,65	70	136	99,8	10,15	99,78	11,15	0,979
Socioemocional	93,05	12,01	65	140	92,13	11,72	94	12,25	0,056
ASQ-3 (adaptado), puntajes crudos									
Resolución de problemas					47,69	15,65			
Comunicación					47,7	18,5			
Motricidad fina					47,2	14,81			
Motricidad gruesa					50,41	16,83			
Personal-social					48,35	14,55			
Denver-II, puntajes crudos									
Lenguaje					20,66	6,39			
Motricidad fina adaptativa					18,97	4,29			
Motricidad gruesa					21,07	5,42			
Personal-social					16,07	4,95			
BDI-2 (Battelle), puntajes crudos									
Desarrollo cognitivo							17,02	3,98	
Comunicación							17,58	6,66	
Motricidad							18,66	6,37	
Personal-social							16,32	5,43	
Habilidades adaptativas							18,79	5,81	
SFI y el SFII (MacArthur), puntajes crudos									
Lenguaje receptivo (SFI)					47,96	21,91			
Lenguaje expresivo (SFI)					6,91	7,94			
Lenguaje expresivo (SFII)					53,85	26,8			
WHO-Motor, puntajes crudos									
Motricidad gruesa							4,07	2,08	
Componentes del FSIQ (WISC-V), puntajes crudos									
Construcción con cubos	15,58	7,32	0	42	15,62	7,44	15,54	7,19	0,855
Semejanzas	12,37	5,92	0	33	12,35	5,87	12,39	5,99	0,912
Matrices	10,01	4,31	0	23	9,99	4,48	10,03	4,13	0,873
Dígitos	14,94	5,63	0	36	14,88	5,73	15,01	5,53	0,733
Claves	30,65	9,94	1	71	31,04	9,97	30,25	9,91	0,223
Vocabulario	14,93	4,72	0	34	14,87	4,75	15	4,69	0,664
Balanzas	12,64	3,81	0	26	12,69	3,72	12,6	3,9	0,714
Componentes del puntaje del desempeño académico, puntajes crudos									
Aritmética, WM-III	8,71	4,23	0	22	8,7	4,28	8,73	4,18	0,917
Comprensión lectora, WM-III	18,38	8,89	0	36	18,55	8,87	18,21	8,91	0,555
Vocabulario receptivo, basado en el TVIP	31,92	13,64	2	72	31,62	13,55	32,23	13,74	0,494

Tabla A2. Correlaciones de los puntajes crudos de los componentes del Coeficiente Intelectual (FSIQ) y del desempeño académico con la edad y el grado escolar en la infancia intermedia

	Edad en meses	Grado escolar
Componentes del FSIQ (WISC-V), puntajes crudos		
Construcción con cubos	0,389***	0,372***
Semejanzas	0,375***	0,408***
Matrices	0,399***	0,422***
Dígitos	0,522***	0,556***
Claves	0,136***	0,211***
Vocabulario	0,388***	0,434***
Balanzas	0,323***	0,353***
Componentes del puntaje del desempeño académico, puntajes crudos		
Aritmética, WM-III	0,592***	0,633***
Comprensión lectora, WM-III	0,621***	0,689***
Vocabulario receptivo, basado en el TVIP	0,502***	0,519***

N = 937; *** p<0,001

Tabla A3. Prueba de significancia (*P*-valores) de las correlaciones entre los grupos etarios: escalas de desarrollo cognitivo y de lenguaje del Bayley-III, las pruebas cortas, el FCI, la talla para la edad y la desnutrición crónica en la primera infancia con el posterior Coeficiente Intelectual (FSIQ) y desempeño académico

	Grupo etario más joven (6 a 18 meses) e intermedio (19 a 30 meses)		**Grupo etario intermedio (19 a 30 meses) y más grande (31 a 42 meses)**		**Grupo etario más joven (6 a 18 meses) y más grande (31 a 42 meses)**	
Pruebas en la primera infancia, 6-42 meses	**FSIQ (CI)**	**Desempeño Académico**	**FSIQ (CI)**	**Desempeño Académico**	**FSIQ (CI)**	**Desempeño Académico**
Bayley-III						
Desarrollo cognitivo	**0,026**	**0,028**	**0,038**	>0,05	**0,000**	**0,000**
Lenguaje receptivo	**0,019**	**0,010**	>0,05	>0,05	**0,000**	**0,001**
Lenguaje expresivo	**0,023**	**0,005**	>0,05	>0,05	**0,000**	**0,000**
ASQ-3 (adaptado)						
Resolución de problemas	**0,003**	>0,05	**0,000**	**0,004**	>0,05	**0,005**
Comunicación	>0,05	>0,05	>0,05	>0,05	>0,05	>0,05
Denver-II						
Lenguaje	>0,05	**0,003**	>0,05	>0,05	**0,001**	**0,000**
Motricidad fina adaptativa	>0,05	>0,05	**0,026**	**0,016**	**0,001**	**0,001**
BDI-2 (Battelle)						
Desarrollo cognitivo	>0,05	>0,05	>0,05	>0,05	>0,05	>0,05
Comunicación	>0,05	**0,028**	>0,05	>0,05	>0,05	**0,030**
SFI y SFII (MacArthur)						
Lenguaje expresivo	**0,036**	**0,027**	-	-	-	-
FCI, talla para la edad, desnutrición crónica						
FCI	**0,014**	**0,006**	>0,05	>0,05	>0,05	>0,05
Talla para la edad	>0,05	>0,05	>0,05	>0,05	>0,05	>0,05
Desnutrición crónica	>0,05	>0,05	>0,05	>0,05	>0,05	>0,05

Notas: Número de observaciones para cada correlación comparada, tal como en la Tabla 3. Grupos etarios "más joven" (6 a 18 meses), "intermedio" (19 a 30 meses) y "más grande" (31 a 42 meses) definidos en función de la edad al inicio del estudio (primera infancia). P-valores <=0,05 en negrita. Los *P*-valores se calcularon mediante métodos *Bootstrap*, estratificando por los estratos del diseño: grupo etario y sector socioeconómico (n =2000 repeticiones). El FCI incluye los materiales y actividades de juego. La desnutrición crónica se define en función de una talla para la edad de -2 desviaciones estándar (DE) por debajo de la mediana establecida por la OMS.