Banco Interamericano de Desarrollo

Adaptación, validación y propiedades psicométricas del ASQ-3 y del Bayley-III en niños menores de 42 meses de zonas rurales de Perú

M. Caridad Araujo
Marta Dormal
Fabiola Lazarte
Beatriz Oré
Marta Rubio-Codina

División de Protección Social y Salud

NOTA TÉCNICA Nº
IDB-TN-1685

Junio 2019

Adaptación, validación y propiedades psicométricas del ASQ-3 y del Bayley-III en niños menores de 42 meses de zonas rurales de Perú

M. Caridad Araujo
Marta Dormal
Fabiola Lazarte
Beatriz Oré
Marta Rubio-Codina

Junio 2019

BID
Banco Interamericano de Desarrollo

Catalogación en la fuente proporcionada por la
Biblioteca Felipe Herrera del
Banco Interamericano de Desarrollo
Adaptación, validación y propiedades psicométricas del ASQ-3 y del Bayley-III en niños menores de 42 meses de zonas rurales de Perú / M. Caridad Araujo, Marta Dormal, Fabiola Lazarte, Beatriz Oré, Marta Rubio-Codina.
p. cm. — (Nota técnica del BID ; 1685)
Incluye referencias bibliográficas.
1. Child development-Peru-Testing. 2. Psychological tests for children-Peru. 3. Rural children-Peru. I. Araujo, M. Caridad. II. Dormal, Marta. III. Lazarte, Fabiola. IV. Oré Luján, Beatriz. V. Rubio-Codina, Marta. VI. Banco Interamericano de Desarrollo. División de Protección Social y Salud. VII. Serie.
IDB-TN-1685

http://www.iadb.org

scl-sph@iadb.org

www.iadb.org/es/proteccionsocial

Adaptación, validación y propiedades psicométricas del ASQ-3 y del Bayley-III en niños menores de 42 meses de zonas rurales de Perú

M. Caridad Araujo[a] , Marta Dormal[a], Fabiola Lazarte[b], Beatriz Ore[c] y Marta Rubio-Codina[a]

Resumen

Un desafío en la evaluación de los servicios de desarrollo infantil es encontrar instrumentos de medición del desarrollo que sean lo suficientemente confiables y válidos para medir sus impactos. Dado el alto costo y la complejidad de administración de pruebas de diagnóstico—consideradas como las más adecuadas para este propósito—cada vez es más común el uso de pruebas de tamizaje como alternativa por ser más baratas, cortas y fáciles de administrar. Sin embargo, la evidencia sobre su validez cuando se administran para una evaluación de impacto, en poblaciones distintas a las poblaciones para las que fueron diseñadas, y en el contexto de encuestas de hogares, es todavía limitada. Este documento describe el proceso de adaptación y las propiedades psicométricas (confiabilidad y validez) de una prueba de tamizaje, el ASQ-3, y de una prueba de diagnóstico, el Bayley-III, ambas administradas como parte de la evaluación de impacto del *Programa Nacional Cuna Más* en Perú. El ASQ-3 mostró una consistencia interna entre baja y moderada, aunque las correlaciones entre las escalas fueron más bajas que las reportadas en el manual. Su validez predictiva fue muy baja. El Bayley-III presentó una consistencia interna aceptable (alfa de Cronbach>0,8) y correlaciones entre las escalas que se acercan a las del manual, excepto para la escala de lenguaje expresivo. Su validez de criterio fue baja, en particular con respecto a variables de calidad del ambiente del hogar. Esos resultados refuerzan la necesidad de investigar alternativas viables y confiables para medir los impactos de programas de desarrollo infantil a escala y en contextos de países en desarrollo.

Códigos JEL: J1, I1, I2, I3

Palabras Claves: adaptación, validación, medición del desarrollo infantil temprano, confiabilidad, validez, ASQ, Bayley, Cuna Más.

[a]Banco Interamericano de Desarrollo, Washington DC, Estados Unidos
[b]Instituto de Investigación Nutricional, Lima, Perú
[c]Universidad Antonio Ruiz de Montoya, Lima, Perú

1. Introducción

Investigaciones provenientes de diversas disciplinas evidencian la importancia de los primeros años de vida para la formación del capital humano (Heckman 2007) y cómo la adversidad en este periodo crítico tiene efectos de largo plazo (Walker et al. 2011; Grantham-McGregor et al. 2007). En América Latina y el Caribe (ALC), las brechas socioeconómicas en el desarrollo se hacen evidentes desde muy temprano y son crecientes con la edad (Berlinski y Schady 2015; Rubio-Codina et al. 2016b; Schady et al. 2015). Los servicios de desarrollo infantil pueden tener efectos positivos en la trayectoria de los niños que los reciben, por ejemplo, en el desempeño escolar (Currie 2009; Cunha et al. 2006) e incluso en los resultados laborales en la edad adulta (Gertler et al. 2014; Heckman, Stixrud, y Urzúa 2006; Maluccio et al. 2009). Esta evidencia ha contribuido a que los servicios de desarrollo infantil dirigidos a poblaciones vulnerables adquieran mayor importancia dentro de la agenda de la política social en ALC y a nivel global.

En 2012, Perú creó el *Programa Nacional Cuna Más* (Cuna Más, en adelante). Pronto se convirtió en el mayor proveedor de servicios de desarrollo infantil del país. Atiende a menores de 36 meses en zonas de pobreza y pobreza extrema. Cuna Más opera a través de dos servicios: el *Servicio de Cuidado Diurno* (SCD), que atiende en centros de cuidado infantil en zonas urbano-marginales; y el *Servicio de Acompañamiento a Familias* (SAF), que opera en comunidades rurales y brinda visitas domiciliaras semanales a niños menores de tres años y sus cuidadores principales. En julio de 2018, el SCD atendía a 56.641 niños y el SAF a 103.063 niños (Cuna Más, comunicación personal, agosto 15, 2018).

Desde su creación, el Gobierno de Perú decidió evaluar ambos servicios: Araujo et al. (2018) documentan la evaluación de impacto del SAF y Guerrero y Léon (2017) la del SCD (con unas muestras de 5.858 y 3.137 niños, respectivamente). Un reto importante en el diseño de estas evaluaciones fue encontrar instrumentos de medición del desarrollo infantil temprano (DIT) que fueran lo suficientemente costo-efectivos para uso en muestras de estos tamaños y en contextos de alta vulnerabilidad y, en el caso de SAF, alta ruralidad y dispersión.

Las pruebas de diagnóstico, tales como las diferentes versiones de las Escalas de Bayley o de las Escalas de Griffiths, miden varios dominios y se consideran las herramientas más adecuadas para la medición del desarrollo (Frongillo et al. 2014; Fernald et al. 2017; Fernandes et al. 2014), en parte porque lo hacen a lo largo de toda la distribución de habilidades. Además, han mostrado ser sensibles a cambios en niveles de desarrollo resultado del impacto de intervenciones en contextos diversos, como Jamaica (Grantham-McGregor 1991), Bangladesh (Nahar et al. 2009) y Colombia (Attanasio et al. 2014), entre otros.

No obstante, la administración de este tipo de pruebas en el contexto de encuestas de hogares como las que se hicieron para la evaluación de Cuna Más es poco factible dado lo costosos que son los materiales, las licencias de uso y los gastos de capacitación, así como por la complejidad de su administración, que requiere de profesionales capacitados y de espacios adecuados. Además, estos instrumentos han sido diseñados en países occidentales y validados en muestras muy distintas a los beneficiarios de SAF. La adaptación de estas pruebas a otros idiomas y contextos culturales es compleja y requiere de expertos, tiempo y presupuesto (Rubio-Codina et al. 2016a)—recursos a

menudo muy limitados en evaluaciones de impacto. Estas restricciones son particularmente relevantes para la medición del DIT en menores de 3 años ya que para este grupo etario existen menos pruebas y la medición del desarrollo es más compleja.

Dado lo anterior, es común el uso de pruebas de tamizaje o de pruebas que miden solamente un dominio del desarrollo (por ejemplo, el lenguaje) como alternativas (Ángeles et al. 2011; Bernal 2015; Fernald e Hidrobo 2011; Macours, Schady y Vakis 2012). Al ser más cortas y recogerse parcial o totalmente por reporte del cuidador, pueden ser más fácilmente administradas por encuestadores no especializados y en el hogar del niño. Por estos motivos, se optó por usar una prueba de tamizaje, los *Cuestionarios de Edades y Etapas* (Ages and Stages Questionnaires, tercera edición, ASQ-3; Squires et al. 2009) para la evaluación de impacto de los servicios de Cuna Más. No obstante, como toda prueba de tamizaje, el ASQ-3 no está diseñado para medir desarrollo a lo largo de toda la distribución de habilidades, sino para detectar niños con rezago o riesgo de rezago en su desarrollo. Además, requería ser adaptado al contexto local.

Rubio-Codina et al. (2016a) analizan la validez concurrente de tres pruebas de tamizaje y dos pruebas que miden un único dominio cuando se administran en el contexto de una encuesta de hogar, por entrevistadores y en una muestra de 1.311 niños de 6 a 42 meses en Bogotá. Las comparan con las Escalas Bayley, administradas en un centro por psicólogos. Los autores encuentran que la validez concurrente aumenta con la edad del niño para las escalas cognitivas, de lenguaje y de motricidad fina y decrece para las de motricidad gruesa. El ASQ-3 es una de las pruebas más cortas y baratas, pero su validez es muy pobre hasta los 30 meses. Otros estudios sobre el tema, en áreas rurales de Bangladesh, exploran la validez concurrente y predictiva de dos pruebas que miden un único dominio: el lenguaje (Hamadani et al. 2010) o la motricidad gruesa (Hamadani et al. 2013). Los autores encontraron niveles de validez concurrente y predictiva moderados con respecto a las Escalas del Bayley y al coeficiente intelectual a los 5 años. No obstante, ninguno de estos estudios investiga la capacidad de estas pruebas de identificar cambios en desarrollo derivados de una intervención.

Para la evaluación del SAF, se decidió complementar el uso del ASQ-3 con la administración de la tercera edición de las Escalas Bayley de Desarrollo Infantil (Bayley 2006; Bayley-III), considerada la prueba de referencia para menores de 42 meses (Fernald et al. 2017). Por las complejidades de adaptación y administración mencionadas, así como las limitaciones de recursos financieros y de tiempo, solo se administró a una submuestra de los niños de la muestra de evaluación y solo en la medición de seguimiento (no en línea de base).

Este estudio describe el proceso de adaptación del ASQ-3 y el Bayley-III en Perú y analiza la confiabilidad y validez de estas pruebas cuando se usan en población vulnerable. El análisis de confiabilidad (consistencia interna) usa el alfa de Cronbach y las correlaciones de las escalas de una misma prueba entre sí. Para el análisis de validez, investigamos la correlación de las pruebas con un conjunto de variables teóricamente relacionadas con el desarrollo infantil (validez de constructo), la correlación entre las escalas coincidentes de ambos instrumentos (validez concurrente), y la correlación entre el ASQ-3 en la línea de base con el Bayley-III en línea de seguimiento (validez predictiva, solo para el ASQ-3).

El estudio está organizado de la siguiente manera. La Sección 2 describe el contexto peruano y las experiencias previas del uso del ASQ y del Bayley en el país. La Sección 3 presenta los datos y las características de las muestras del estudio. La Sección 4 describe las dos pruebas, su proceso de adaptación al contexto peruano y la metodología de

análisis. La Sección 5 expone los resultados y la Sección 6 provee una discusión sobre los mismos.

2. Contexto y Antecedentes

2.1 Perú

Perú es un país de ingreso medio que en los últimos 20 años ha experimentado una fuerte expansión económica. Con una tasa de crecimiento promedio por encima del 5% (BID 2017), ha logrado reducir los niveles de pobreza. En la última década, la proporción de la población debajo de la línea de pobreza del Banco Mundial de 3,10 dólares per cápita se redujo del 27 al 9% (Banco Mundial 2017a). También se observaron mejoras en varios indicadores sociales. En 2015, la prevalencia de la desnutrición crónica para los menores de 5 años se redujo a 14% (la mitad de lo que era durante la década anterior) y la tasa de mortalidad infantil bajó a 15 por mil nacidos vivos (Banco Mundial 2017b). En cuanto a la educación, las tasas de matrícula en preescolar (3-5 años) incrementaron del 53% en 2001 al 82% en 2015.

Sin embargo, Perú es un país que muestra grandes contrastes entre las zonas urbanas y rurales, entre las regiones andina y amazónica, y entre grupos socioeconómicos. A pesar de que menos del 25% de la población vive en el ámbito rural, cerca de la mitad de la pobreza moderada y el 80% de la pobreza extrema se encuentra en estas zonas (Banco Mundial 2016b). La desnutrición crónica en el 2013 era 3 veces más alta en zonas rurales que urbanas, y cerca de 10 veces más frecuente para los niños con madres que no tenían educación formal, en comparación con aquellos cuyas madres tenían educación terciaria (INEI 2017). Cuando ingresan a la escuela primaria, los niños del cuartil de riqueza más pobre en zonas rurales tienen un desempeño de lenguaje 0,77 desviaciones estándar menor que los niños del cuartil de riqueza más alto (Schady et al. 2015). La dispersión y marginalidad de la población rural también representa un desafío para la provisión de servicios públicos. En 2015, a nivel nacional, el 86% de hogares tenía acceso a servicios de agua y el 77% a servicios de saneamiento. Estas estadísticas bajan al 66 y 44% para el área rural (Banco Mundial 2016b). La gran diversidad cultural y multilingüismo del país complejiza todavía más la provisión de servicios y el diseño de políticas públicas.

2.2 Uso del ASQ y del Bayley en Perú

En esta sección revisamos estudios anteriores que administraron el ASQ o las Escalas de Bayley en Perú, en cualquiera de sus ediciones. Identificamos dos limitaciones importantes en estas experiencias, que tratamos de mitigar en el presente estudio. La primera es la falta de una adaptación de las pruebas al contexto de administración. Esto puede causar interpretaciones erróneas de los resultados si, por ejemplo, algunos ítems de la versión original de la prueba no son pertinentes en el contexto local. La segunda es el uso de una estandarización *externa* de los puntajes, es decir, empleando las normas de las poblaciones de referencia para las cuales se desarrollaron los instrumentos (Estados Unidos, en el caso del ASQ y del Bayley), en lugar de una estandarización *interna,* usando las medias y desviación estándar de la muestra. Los niños pueden desarrollarse a ritmos diferentes y, por este motivo, los puntos de corte y normas establecidos por la prueba en un país pueden no ser apropiados en otros contextos.

ASQ. Para el ASQ, Kyerematen et al. (2014) administraron el ASQ-2 (Bricker et al. 1999)

y el ASQ-3 a 129 niños de entre 3 meses y 5 años en la periferia de Lima. El objetivo del estudio era identificar niños con algún retraso en el desarrollo e investigar asociaciones con indicadores nutricionales. Se administraron las cinco escalas—resolución de problemas, comunicación, motricidad fina, motricidad gruesa, y personal-social—y se categorizó como resultado 'sospechoso' cuando el puntaje de un niño se ubicaba debajo del corte establecido por el instrumento para la población de Estados Unidos. El estudio no menciona haber hecho ninguna adaptación de la prueba antes de su administración, a pesar de identificar ítems que lo hubieran ameritado.

Westgard y Alnasser (2017) también tenían como objetivo identificar niños con algún retraso en el desarrollo y explorar sus determinantes sociales. Administraron el ASQ-3 y un cuestionario de hogar a una muestra de 596 niños de entre 8 y 38 meses en la región amazónica. Un ejercicio de validación previo en campo identificó que tanto el lenguaje como los ejemplos presentados eran culturalmente relevantes al contexto cuando estaban acompañados de una explicación apropiada por parte del evaluador. Para reducir el tiempo de administración, decidieron emplear solamente las escalas de comunicación y motricidad gruesa. Al igual que Kyerematen et al. (2014), compararon los puntajes con los puntos de corte establecidos por la prueba.

Fernald et al. (2012) usaron el ASQ para evaluar un proyecto de agua y saneamiento en Perú. Administraron las escalas de comunicación, motricidad gruesa y personal-social a 8.727 niños de entre 3 y 23 meses. Adaptaron la segunda edición del ASQ sustancialmente y crearon el *Extended Ages and Stages Questionnaire* (el ASQ "extendido"; EASQ en adelante) para lo que incluyeron, en cada dimensión, algunos ítems adicionales del grupo etario subsiguiente. También modificaron ítems para reflejar mejor el contexto local, por ejemplo, con respecto a la vestimenta, la comida, los muebles, y los juegos de los niños. Otros ítems fueron descartados por completo si es que se encontraron tasas de respuesta muy bajas. Además, para que la prueba fuera lo más uniforme posible (en cuatro países que formaron parte de este estudio), los autores agregaron la opción "no aplica" a ítems que no eran relevantes en alguno de los países. En cuanto a la estandarización de los puntajes totales, en lugar de los cortes establecidos por el manual, usaron las medias y desviaciones estándar específicas a cada edad.

Nelson et al. (2018) también usaron el EASQ en el contexto de la evaluación de impacto de un programa de estimulación temprana en Lima. Administraron la prueba a una muestra de 60 niños y estandarizaron los puntajes internamente usando las medias y desviaciones estándar específicas a cada edad. No mencionan haber hecho modificaciones adicionales a la prueba con respecto a la versión en Fernald et al. (2012).

Bayley. Las Escalas del Bayley han sido utilizadas en Perú en sus diferentes versiones desde los años sesenta, empezando por el estudio de Pollit y Granoff (1967) que administró las escalas de desarrollo mental y de desarrollo motor a una muestra de 27 niños de entre 11 y 32 meses en Lima, quienes habían sido hospitalizados y tratados por malnutrición. Este, al igual que todos los estudios revisados, usó los puntajes compuestos de la prueba (media 100 y desviación estándar 15), cuyas normas están basadas en la población de Estados Unidos (estandarización externa).

El Instituto de Investigación Nutricional (IIN) ha utilizado esta escala ampliamente en varios estudios que analizan la asociación entre desarrollo infantil y nutrición. Por ejemplo, Oré, Diaz y Penny (2011) usaron la segunda edición de las escalas Bayley (Bayley-II) para una evaluación pre-post de una intervención grupal con 163 díadas madre-hijo en

un distrito urbano-marginal de Lima. Los autores no mencionan haber modificado la prueba.

Por su parte, Colombo et al. (2014) usaron el Bayley-II para evaluar el efecto de la suplementación con zinc sobre el desarrollo sensoriomotor en niños de entre 6 y 18 meses de una zona urbano marginal en Lima. Para este estudio, todos los protocolos de la prueba fueron traducidos y piloteados y se probó que la prueba fuera adecuada al contexto y población de estudio. No se describe ninguna adaptación del instrumento.

Joseph et al. (2015) evaluaron el beneficio, el momento y la frecuencia óptima de la desparasitación sobre el desarrollo infantil en una muestra de 1.563 niños de 12 a 24 meses de edad en Iquitos. El Bayley-III pasó por un proceso de traducción y adecuación en el cual se realizaron dos tipos de modificaciones: primero en la forma de administración (e.g., se especificó, para cada ítem, la cantidad de veces que podía ser repetida la instrucción); y segundo, en la instrucciones verbales e imágenes, que fueron adecuadas al contexto local. Como se detalla más adelante, el presente estudio siguió un enfoque parecido en algunos de estos aspectos. Esta misma versión adaptada se utilizó en el estudio que realizó Blouin et al. (2018) para evaluar el efecto de ser infectado con helmintiasis sobre las habilidades cognitivas y de lenguaje hasta los 3 años en un subgrupo de 880 niños del estudio de Joseph et al. (2015).

Por último, Scharf et al. (2018) administraron el Bayley-III en el departamento de Loreto a menores de 24 meses como parte del estudio MAL-ED (MAL-ED Network Investigators 2014), realizado en 8 países para investigar la asociación entre infecciones entéricas, malnutrición, fisiología intestinal, crecimiento físico, desarrollo cognitivo y respuestas inmunes en niños de escasos recursos. El Bayley-III fue traducido, adaptado y piloteado en cada uno de los países. La prueba pasó por un proceso de traducción inversa (‘back translation’ en inglés) y se adaptaron y pilotearon ítems que requerían una adecuación cultural. También se hicieron análisis psicométricos y se descartaron aquellos ítems que no tenían suficiente variabilidad.

3. Datos

3.1. Recolección de datos

Los datos empleados provienen de la muestra de la evaluación de impacto del SAF. Esta constaba de 5.858 niños de 1 a 24 meses en 180 distritos elegibles para recibir el Programa, repartidos en 12 departamentos del país, principalmente en las regiones andina y amazónica.[1] La evaluación incluyó una encuesta de línea de base (LB en adelante) en 2013 y una encuesta de línea de seguimiento (LS en adelante), aproximadamente 24 meses después, en la segunda mitad de 2015. El ASQ-3 se administró a todos los niños tanto en LB como en LS. En la LS se administró, además, el Bayley-III a una submuestra de 1.492 niños. Esta submuestra difiere en dos características importantes. Primero, tiene una proporción casi nula de niños de hogares cuya lengua materna no es el castellano, dadas las dificultades de traducir la prueba al quechua u otras lenguas indígenas. Segundo, la submuestra está compuesta por los

[1] DGSE-MIDIS et al. (2015) explica en detalle el diseño de la intervención, y los pasos y criterios que llevaron a la selección de esta muestra a partir del marco muestral.

niños más pequeños de la muestra de evaluación (menores de 43 meses en la LS), ya que el rango de administración del Bayley-III es hasta los 42 meses de edad.

3.2. Muestras de análisis

Este estudio emplea dos muestras de análisis, a las que nos referimos como 'muestra ASQ-3' (N= 1.493) y 'muestra Bayley-III' (N=385), de aquí en adelante.

La muestra ASQ-3 se seleccionó de la siguiente manera. De los 5.858 niños en la muestra de evaluación, decidimos incluir solo a aquellos en el grupo de control, de tal manera que nuestra muestra de análisis estuviera 'limpia' del impacto de la intervención (N=1.997). Por comparabilidad con la muestra Bayley-III, excluimos también a aquellos niños que no pertenecen a hogares indígenas, usando como criterio el indicador de si la madre contestó la encuesta de la LB en un idioma indígena (N=1.686). Finalmente, limitamos la muestra a los niños que fueron medidos con el ASQ-3 tanto en la LB como en la LS (N=1.493).

La muestra Bayley-III es la submuestra de los 385 niños de la muestra ASQ-3 (N=1.493) a quienes se les administró el Bayley-III en LS.[2]

3.3. Características de los niños, de las familias, y de los hogares

La Tabla 1 presenta, para ambas muestras, las características de los niños, sus familias y hogares en la LB. Los indicadores de desnutrición crónica y global presentados se calcularon usando los lineamientos de la Organización Mundial de la Salud (WHO Multicentre Growth Reference Study Group 2006).[3]

Para caracterizar la calidad del ambiente del hogar, reportamos algunos de los Indicadores del Cuidado Familiar o *Family Care Indicators* (FCI; Kariger et al. 2012): el número de juguetes con los que el niño juega usualmente; el número de actividades de juego que el niño realizó con una persona mayor de 15 años en la semana previa a la entrevista; y si el hogar cuenta con libros para adultos. Comparamos los promedios para cada variable e incluimos el p-valor de la hipótesis nula que la diferencia entre ellas es igual a cero.

[Tabla 1]

Existen algunas diferencias significativas entre las dos muestras. Por diseño, los niños de la muestra Bayley-III son menores que los niños de la muestra ASQ-3 (*M*=7,54 y *M*=12,81, respectivamente; p-valor=0,00). Otras de las diferencias existentes podrían estar relacionadas con la composición socioeconómica de cada muestra, como resultado de los criterios establecidos para la definición de la muestra Bayley-III. Por ejemplo, los niños en la muestra Bayley-III tienen una tasa de desnutrición crónica menor (31% vs. 36%, p-valor=0,01) y la muestra tiene una proporción más alta de niños en la sierra urbana (26% vs. 20%, p-valor=0,00) y una proporción menor en la selva (17% vs. 23%, p-valor=0,00).

Las dos muestras tienen tasas de desnutrición global elevadas (alrededor del 5-7%). La media del peso al nacer es adecuada en ambas muestras y las tasas de prematuridad se encuentran en alrededor del 7-8%. El 29-30% de los niños son primogénitos y el 73% recibieron lactancia materna exclusiva durante 6 meses o más.

[2] Esto excluye también a 18 niños mayores de 42 meses cuando se administró el Bayley-III.
[3] http://www.who.int/childgrowth/es/

Las características de las familias son parecidas entre las dos muestras: en promedio, las madres tienen 28 años y alrededor de 7 años de escolaridad, y los padres 8 años de educación.[4] El 36-37% de las madres trabaja y, aunque la mayoría de ellas habla el castellano, el 31-32% reporta una lengua indígena como lengua materna.[5]

En lo que se refiere a la tenencia de activos y al acceso a servicios básicos, se trata de hogares vulnerables en ambas muestras: solo 6-8% tiene refrigerador, 40-41% cocina a gas y 27-28% de viviendas cuenta con un piso que no es de tierra. Un 57-59% de las viviendas tiene acceso a agua y 22-25% posee un servicio higiénico dentro de la vivienda, en ambas muestras. Una proporción un poco más alta de hogares en la muestra Bayley-III tiene televisor (51% vs. 47%, p-valor=0,06) y celular (70% vs. 63%, p-valor=0,00).

Los niños no cuentan con gran variedad de materiales de juego en casa ni realizan muchas actividades de juego con adultos. De los ocho tipos de juguetes/materiales considerados, los niños en la muestra ASQ-3 tienen en promedio 1,67 variedades vs. 1,31 variedades en la muestra Bayley-III (p-valor=0,00). Con la excepción de los libros infantiles de cuentos y las muñecas y objetos para el juego de roles, los niños en la muestra de Bayley-III tienen una menor cantidad de todos los juguetes analizados. En parte, esto puede ser debido a su menor edad (*M*=12,81 para la muestra del ASQ-3, y *M*=7,54 para la muestra del Bayley-III).

Respecto a las actividades de juego, se encuentra que los niños realizaron en promedio entre dos y tres actividades (de las siete evaluadas) durante la semana previa a la encuesta. Se observa una media más alta en la muestra ASQ-3 (2,86 vs. 2,46, p-valor=0,00), que proviene de una menor frecuencia en todas las actividades evaluadas en la muestra Bayley-III (con una sola excepción, jugar con los juguetes de los niños). Nuevamente, esto puede explicarse por la diferencia de edad entre los niños de ambas muestras.

4. Métodos

4.1. Instrumentos y adaptación

A continuación, se describen las características generales y procesos de adaptación del ASQ-3 y del Bayley-III.

4.1.1. El Cuestionario de Edades y Etapas, tercera edición (ASQ-3)

El ASQ-3 (Squires et al. 2009) es una prueba de tamizaje para niños de 1 a 66 meses de edad. Está compuesta por 21 cuestionarios específicos a la edad que debe responder el cuidador principal. Cada cuestionario evalúa el desarrollo del niño en cinco escalas: resolución de problemas (o cognición), comunicación, motricidad fina, motricidad gruesa y personal-social. Cada escala contiene seis ítems. Cada ítem se responde con las alternativas 'sí' (10 puntos), 'a veces' (5 puntos) o 'todavía no' (0 puntos). Para cada

[4] El 9% de ellas no tiene educación, el 19-22% tiene la educación primaria completa, el 19-20% la secundaria completa, y el 6-8% siguieron estudiando después de la secundaria.
[5] Si bien la muestra excluye aquellas madres que contestaron la encuesta de la LB en una lengua distinta al castellano, esta pregunta se refiere a la lengua que la madre considera como materna.

escala, se suman los puntajes de cada ítem, teniendo un puntaje máximo de 60 puntos por escala.

El nivel de precisión de esta prueba en la medición de las capacidades reales del niño en cada escala del desarrollo es limitado, en particular para niños con un desempeño ubicado en el rango normal o por encima de éste.

Los autores de la prueba[6] señalan que el ASQ-3 puede utilizarse adecuadamente con familias y niños de diversas culturas y bagajes lingüísticos. Para ello sugieren realizar algunas adaptaciones, las cuales sirvieron de base para la adaptación del ASQ-3 al contexto rural peruano.

La adaptación de los cuestionarios se llevó a cabo en dos momentos diferentes, pero siguiendo el mismo procedimiento. El primer trimestre de 2013 se adaptaron los cuestionarios correspondientes a los niños menores de 25 meses para administración en la LB; y, dos años después, en el primer trimestre de 2015, se hicieron las adaptaciones para los que tenían entre 25 y 54 meses para administración en la LS. La adaptación siguió seis fases:

1. Consulta: se realizaron consultas de inspiración etnográfica a través de entrevistas a conocedores de la cultura local y grupos focales con madres y otros actores locales (e.g., personal de los servicios de educación y salud).

2. Adecuación: se realizaron las siguientes adecuaciones a la versión original del instrumento. Primero, con respecto a la forma de administración, se decidió que los ítems con mayor complejidad serían evaluados mediante la observación directa del niño por parte del encuestador, en lugar de recoger la información por reporte del cuidador principal. En zonas rurales del Perú, el grado de instrucción de la madre es muy básico[7] y tal como se señala en la guía de adaptación del ASQ[8], algunas madres creen que son los especialistas quienes mejor conocen el desarrollo de sus niños y que deberían 'evaluarlos'. Es probable que para los ítems del ASQ-3 que evalúan habilidades más complejas, se dificulte la tarea de la madre de identificar el desempeño de sus hijos y contestar a los ítems del cuestionario. En base a esto, el equipo técnico a cargo de la adecuación del instrumento elaboró una primera selección de ítems que debían ser administrados por la evaluadora, que se validó y ajustó en pilotos. En la versión final del manual de los encuestadores y también en el formulario de la encuesta, se especificó qué ítems tenían que ser administrados y cuáles reportados por el cuidador.

Segundo, se realizaron cambios al contenido de algunos ítems para garantizar la idoneidad cultural de las palabras y frases. Por ejemplo, se reemplazó 'voy al parque' por 'voy a la tienda', dado que en algunas zonas rurales del Perú no existen parques; o 'tenedor' por 'cuchara', ya que en las mismas zonas no es usual comer con tenedor. También se cambiaron algunas palabras utilizando sinónimos o palabras de uso común en el medio rural andino. Otros cambios lingüísticos en los ítems incluyeron la edición de las preguntas para simplificarlas y hacerlas más directas, además de colocar entre paréntesis algunas ideas para que se comprendieran mejor. Para facilitar el trabajo del encuestador, se incluyó un título a cada ítem (e.g., 'nombra una figura'). Así, el encuestador podía identificar rápidamente la habilidad que se estaba evaluando y el concepto general del ítem. También se decidió incluir el nombre del niño en las preguntas

[6] https://www.brookespublishing.com/product/asq-3/
[7] El promedio en ambas muestras del estudio es 7 años de educación.
[8] http://archive.brookespublishing.com/documents/ASQ-cultural-and-linguistic-adaptation.pdf

y empezar cada ítem por ‘Señora’, para favorecer una interacción cercana y respetuosa entre la madre y el encuestador.

En base a adaptaciones en estudios anteriores (Rubio-Codina et al. 2016a; Fernald et al. 2012), y con la finalidad de ampliar la capacidad de medición de la prueba en niños que se ubican en el extremo superior de la distribución de puntajes, se incrementó el número de ítems a ser administrados por cuestionario. Específicamente, si el niño lograba la puntuación máxima en una escala (60 puntos), se evaluaban adicionalmente los primeros tres ítems nuevos (es decir, no coincidentes) del cuestionario subsiguiente. La Tabla A1 en el anexo muestra la diferencia en el porcentaje de niños que alcanzan el puntaje máximo en las versiones con 6 y con 9 ítems. En la LB, este porcentaje disminuyó del 15-20% en la versión con 6 ítems al 3-5% en la versión con 9 ítems; y en la LS, del 4-23% al 0-10%.

3. Pre-piloto y ajustes: se realizaron pruebas en zonas rurales (Iquitos, Cusco, La Libertad y Lima) para analizar el desempeño de las versiones adaptadas del ASQ-3. Luego se analizaron los resultados y se incorporaron cambios adicionales derivados de la experiencia de campo.

4. Discusión de gabinete: un grupo de trabajo conformado por psicólogas especialistas en desarrollo infantil y en administración de encuestas revisó los cuestionarios y propuso modificaciones.

5. Capacitación: el perfil de quienes fueron capacitados incluía personas con estudios técnicos en educación, enfermería, obstetricia o nutrición; y experiencia en trabajo de campo aplicando encuestas y manejo de la lengua nativa (quechua) de al menos una de las zonas en donde se realizaría el estudio.

La capacitación de la LB fue de 14 días y la metodología incluyó sesiones teóricas, entrevistas simuladas en aula con madres y niños, plenarias para uniformizar criterios técnicos y/o resolver dudas, así como prácticas de campo evaluadas. Se capacitó a 110 encuestadores y se seleccionó a los 76 con el mejor desempeño en la capacitación y prácticas. Además, se contó con la participación de 12 coordinadores departamentales y 45 supervisores locales con experiencia en encuestas de hogares.

En la LS se trabajó con un equipo más pequeño, de 20 encuestadores, capacitados en 13 días. Esto permitió acompañar más de cerca las sesiones prácticas.

Las psicólogas que trabajaron en la adecuación del instrumento estuvieron a cargo de ambos procesos de capacitación (LB y LS).

6. Piloto y Ajustes: luego de capacitar los encuestadores en el uso de la versión adaptada del ASQ-3, se realizó una administración piloto en zonas rurales de Lima. Se recogieron los comentarios y sugerencias de los encuestadores para realizar ajustes a los cuestionarios y tener la versión final que se utilizó en la encuesta.

4.1.2. Escalas de Desarrollo Infantil de Bayley, tercera edición (Bayley-III)

El Bayley-III (Bayley 2006) es una prueba de diagnóstico que evalúa varias dimensiones del desarrollo: cognitivo, lenguaje receptivo, lenguaje expresivo, motricidad fina, motricidad gruesa, socioemocional y comportamiento adaptativo. Cada dimensión se administra y puntúa de forma independiente. Las áreas cognitiva, de lenguaje y motoras se evalúan a través de la observación directa de las habilidades del niño en varios ítems en orden ascendente de dificultad. Existen criterios de inicio y parada, que determinan los

ítems de la prueba que realiza cada niño. Por cada ítem que el niño realiza correctamente recibe un puntaje de 1 (o de 0, si no logra ejecutarlo). El puntaje bruto es la suma de respuestas correctas, incluyendo los ítems previos al punto de inicio. La escala motora gruesa es logísticamente más compleja porque requiere de más materiales (e.g., escalones y cintas) y más espacio. Dadas las restricciones de tiempo y presupuestarias, se administraron exclusivamente las escalas de desarrollo cognitivo, lenguaje (expresivo y receptivo), y motora fina.

Como se mencionó en la Sección 2, en Perú se han utilizado diferentes versiones de la prueba y con adecuaciones para el contexto amazónico (Joseph et al. 2015). La adaptación del Bayley-III que se usó para la evaluación del SAF se hizo a partir de la versión en español empleada previamente en estudios en Colombia (Attanasio et al. 2014; Rubio-Codina et al. 2016a) y en Perú (Colombo et al. 2014; Joseph et al. 2015). El proceso fue liderado por el mismo grupo de psicólogas que realizó las adaptaciones al ASQ-3. Este proceso fue importante para evitar los sesgos que se pueden producir al utilizar una prueba elaborada para otra cultura y población. Por este motivo, el equipo se concentró en asegurar la equivalencia lingüística y funcional entre la versión original y la versión adaptada. Específicamente, el proceso de adaptación de los cuestionarios del Bayley-III siguió las siguientes cuatro fases:

1. Versión preliminar: esta fase incluyó (i) la revisión de las versiones anteriores de la prueba que habían sido usadas en contextos similares y la elaboración de propuestas de adaptación; y (ii) el desarrollo de un manual preliminar y protocolos de administración para cada ítem, incluyendo el máximo número de veces que las instrucciones se podían repetir al niño. En el diseño de los protocolos se prestó particular consideración al hecho que la prueba se iba a administrar, por primera vez, en el contexto andino en Perú. El equipo se concentró también en desarrollar instrucciones para la administración de los ítems.

2. Pre-piloto y ajustes: en esta fase se llevó a cabo la administración de las pruebas por parte de las dos psicólogas encargadas de la adecuación a 22 niños de 24 a 42 meses en dos distritos rurales de Cuzco y Cajamarca. Durante estas administraciones, las psicólogas registraron la equivalencia funcional, lingüística y conceptual de los ítems, de las instrucciones y de las imágenes. También se hicieron grupos focales con las madres y el personal local de los servicios de educación y de salud para recoger sugerencias sobre los ítems y las imágenes. Como resultado de este proceso, se ajustaron algunos de ellos: por ejemplo, para la escala de lenguaje receptivo, se cambiaron las imágenes de niños en una piscina por niños bañándose en un caño; de una pelota de básquet por una de fútbol; y de una ardilla por un cuy.

3. Capacitación: se entrevistó a 16 psicólogas profesionales con experiencia previa en evaluación de niños, de entre las que se seleccionaron 10 evaluadoras y 2 supervisoras, quienes realizarían la coordinación con las comunidades para la evaluación, encontrarían el espacio para la administración de la prueba, explicarían a las familias el objetivo del estudio y recogerían los consentimientos informados respectivos. Las 12 personas seleccionadas fueron entrenadas durante 6 semanas, que incluyeron sesiones teóricas y prácticas. Las sesiones teóricas cubrían la revisión, discusión y demostración de los ítems, ya sea a través de videos o por las mismas capacitadoras. Además, se realizaron prácticas con niños en un centro de salud de Villa el Salvador, una zona urbano-marginal de Lima, durante dos semanas. Se escogió un distrito urbano-marginal de manera que el personal capacitado tuviera la experiencia y la práctica

de evaluar a niños de la misma edad de la muestra del estudio y de características socioeconómicas lo más similares posibles. Además, se consideró de suma importancia que estas prácticas se realizaran en las condiciones óptimas de evaluación descritas en el manual de la prueba. Las evaluadoras se agruparon en parejas, de manera que cada pareja participaba en una evaluación. Cada evaluadora logró administrar la prueba de 10 a 12 veces. De este total, las capacitadoras lograron acompañar entre 5 y 7 administraciones de cada evaluadora. Esto permitió dar una retroalimentación individual al desempeño durante la práctica y lograr medir el índice de confiabilidad intersujeto en cada administración (a partir del cuarto día de práctica), el cual se estableció que hacia el final de las prácticas debía alcanzar un 90% de concordancia como mínimo. Durante esta práctica, se logró administrar la prueba a un total de 113 niños de entre 24 y 42 meses. Una de las 10 evaluadoras que inició la capacitación se retiró durante las prácticas. De las 9 restantes, se seleccionaron 8 de acuerdo a su puntaje en las confiabilidades intersujeto (medido por el porcentaje de acuerdos) y su tasa de asistencia a la capacitación.

4. Piloto y ajustes: El piloto se realizó durante 5 días en Huarochirí, provincia ubicada en la sierra de Lima. Las evaluaciones se llevaron a cabo en los centros comunitarios, en espacios de las municipalidades o centros de estimulación temprana. Fueron evaluados un total de 59 niños; 33 de 24 a 36 meses y 26 de 37 a 42 meses**.** Cada evaluadora pudo evaluar entre 5 y 6 niños en total, en varias comunidades. Las dos capacitadoras pudieron acompañar entre 2 y 3 administraciones de cada evaluadora durante el piloto. En cada administración que acompañaban, las capacitadoras daban una retroalimentación a la evaluadora, calculaban la confiabilidad intersujeto y calificaban a la evaluadora según criterios previamente establecidos. En estas prácticas acompañadas, se logró llegar hasta niveles de confiabilidad de más de 90% con cada una de las evaluadoras.

La versión final de la prueba adaptada pasó por el proceso de traducción inversa con un traductor oficial con experiencia en temas de DIT.

En esta fase, el Comité Institucional de Ética de la Universidad Peruana Cayetano Heredia revisó los protocolos de estudio y los consideró acorde con las prácticas éticas requeridas. Los padres de los niños participantes firmaron el consentimiento informado.

4.2. Análisis estadístico

Presentamos un análisis descriptivo de los puntajes de ambas pruebas (Tabla 2).

Después, exploramos la confiabilidad de los instrumentos, empezando por la consistencia interna de las pruebas: Tabla 3 para el ASQ-3 y Tabla 4 para el Bayley-III. La confiabilidad de la consistencia interna examina hasta qué punto los ítems de una escala (o prueba) miden el mismo constructo subyacente (habilidad). Esta medida de confiabilidad se puede estimar calculando el alfa de Cronbach (α) en todos los ítems de la escala. Los índices más altos de confiabilidad indican un mejor desempeño de la prueba en una población determinada. El análisis de confiabilidad es particularmente importante cuando una prueba se administra en una población diferente de aquella para la cual fue desarrollada, en especial si el idioma o el contenido de los ítems han sido modificados a fin de asegurar su comprensión y equivalencia lingüística. Dado que el ASQ-3 tiene cuestionarios que son específicos a la edad, calculamos primero las alfas de Cronbach por cuestionario (i.e., se le asigna un alfa a cada niño según el cuestionario que le fue administrado) y

después calculamos el promedio de las alfas para los niños en cada grupo etario. También comparamos los resultados para los puntajes con 9 y con 6 ítems. Además, analizamos en qué medida las escalas de una misma prueba se correlacionan entre sí calculando un coeficiente de correlación de Pearson. Este es un indicador del grado de congruencia que se establece entre las escalas y de la interrelación que existe entre ellas (Tabla 5).

Después, presentamos los resultados de un análisis de validez, esto es, en qué medida la prueba mide lo que se supone debe medir. En este estudio, nos concentramos en la validez de criterio—es decir, en la correlación de los resultados de las pruebas con otras variables con las cuales se esperaría que estuvieran correlacionados. Empezamos analizando la validez concurrente entre las puntuaciones en cada escala de las dos pruebas y un conjunto de variables teóricamente relacionadas con el desarrollo infantil mediante el cálculo del coeficiente de correlación de Pearson (Tabla 6). Las variables son las reportadas en la Tabla 1. Para el estatus nutricional, usamos los indicadores de talla por edad y peso por edad estandarizados internamente con media cero y desviación estándar uno (puntajes z). Para los materiales y las actividades de juego del FCI, construimos dos índices agregando los ítems de cada conjunto de preguntas mediante análisis factorial y expresamos los totales también como puntajes z. Además, construimos un índice de riqueza por análisis de componentes principales con la información sobre los activos y el acceso a servicios de las familias.

También presentamos los resultados de la validez concurrente del ASQ-3 y del Bayley-III, ambos medidos en LS, mediante correlaciones de Pearson (Tabla 7). Finalmente, exploramos la validez predictiva del ASQ-3 presentando las correlaciones de Pearson entre el ASQ-3 en LB con las del ASQ-3 en LS, y entre el ASQ-3 en LB y el Bayley-III en LS (Tabla 8). La validez concurrente y predictiva (Tablas 7 y 8) se analizan solo con los niños que están en ambas muestras.

Usamos dos tipos de puntajes en función del análisis. Para la parte descriptiva y el análisis de consistencia interna, empleamos los puntajes crudos para ambas pruebas. Estos se construyen siguiendo las instrucciones de los manuales de administración. Para el ASQ-3, usamos el puntaje que incluye 9 ítems por dos motivos: (i) permite reducir el número de niños que logran el puntaje máximo y consecuentemente aumentar la variabilidad de los puntajes (ver Tabla A1 en el apéndice); y (ii) este puntaje demostró tener una mejor consistencia interna que el puntaje con 6 ítems (ver Tabla 3), aunque posiblemente esto sea también debido al mayor número de ítems. Para el Bayley-III, por 'comparabilidad' con otros estudios, además del puntaje crudo, hemos incluido también los puntajes compuestos. Éstos corresponden a una función no lineal de los puntajes brutos, con una media de 100 y una desviación estándar de 15, en base a las normas establecidas para la población de referencia. Las escalas de lenguaje receptivo y expresivo se combinan para generar un puntaje único de lenguaje. No podemos construir el puntaje compuesto motor porque no administramos la escala de motricidad gruesa, la cual se combina con la de motricidad fina para construir dicho puntaje. Los puntajes brutos del Bayley-III son crecientes con la edad, mientras que los del ASQ-3 no lo deberían ser dado que provienen de cuestionarios específicos para cada edad. Sin embargo, en la práctica se observa que aumentan con la edad.

Para el resto del análisis, y por los motivos explicados anteriormente, usamos los puntajes estandarizados *internamente*, usando la media y desviación estándar específicas por día de edad en la muestra de forma no paramétrica (Rubio-Codina et al. 2016b). Estandarizar los puntajes por este método corrige por el que los puntajes brutos incrementan con la

edad, permitiendo así comparar puntajes entre niños de diferentes edades. A fin de controlar por los efectos relacionados con la idiosincrasia del encuestador sobre la administración y puntuación de la prueba, estandarizamos internamente los *residuos* de los puntajes brutos por edad, netos del efecto de encuestador, en lugar de estandarizar de forma directa los puntajes brutos. Este puntaje se usa en todos los análisis con correlaciones de Pearson[9].

Para facilitar la presentación de los resultados, clasificamos las correlaciones de Pearson como muy bajas cuando r<0,20; bajas en el rango r=[0,20-0,39]; moderadas en el rango r=[0,40-0,59] y altas en el rango r=[0,60-0,79] (Evans 1996). Para el alfa de Cronbach, consideramos un valor de 0,7 como un mínimo deseado.

5. Resultados

5.1. Puntajes del ASQ-3 y del Bayley-III

La Tabla 2 presenta los puntajes crudos del ASQ-3 para ambas muestras y los puntajes crudos y compuestos para la muestra Bayley-III. Para cada escala del ASQ-3, incluimos el p-valor de la hipótesis nula que la diferencia entre las dos muestras es igual a cero.

[Tabla 2]

En cuanto a los puntajes del ASQ-3 en LB, observamos diferencias significativas en tres escalas: comunicación, motora fina y motora gruesa. En dos de ellas, la muestra ASQ-3 tiene puntajes más bajos que la muestra Bayley-III: comunicación (M=43,93 vs. M=49,92; p-valor=0,00) y motricidad fina (M=47,45 vs. M=50,70; p-valor=0,00). En cambio, los niños de la muestra ASQ-3 obtuvieron un puntaje más alto en la escala de motricidad gruesa que los de la muestra Bayley-III (M=43,98 vs. M=41,30; p-valor=0,00).

Con todas las reservas asociadas a usar normas externas en esta población, los puntajes compuestos del Bayley-III indican que los niños están 0,87 desviaciones estándar por debajo de la media de la población de referencia para desarrollo cognitivo y 0,74 desviaciones estándar para lenguaje.

5.2 Análisis de confiabilidad

5.2.1. Consistencia interna

Las Tablas 3 y 4 muestran la consistencia interna (alfa de Cronbach) del ASQ-3 y del Bayley-III para toda la muestra y por grupo etario. La Tabla 3 compara las versiones del ASQ-3 con 6 y 9 ítems.

[Tabla 3]

Para toda la muestra, los coeficientes calculados con los 6 primeros ítems tienen valores inferiores al punto de corte deseado de α≥0,70 (α =[0,26-0,55]). La consistencia interna de la versión con 9 ítems es superior a la versión con 6 ítems y presenta valores entre α=[0,60-0,70], salvo para las escalas de resolución de problemas y motricidad fina en la LS.

Para este análisis también dividimos las muestras en grupos etarios, agrupando a los niños de acuerdo con las edades planteadas en los cuestionarios del ASQ-3 y

[9] Debido a que todas las correlaciones utilizan estos puntajes estandarizados, esto equivale a calcular las correlaciones parciales controlando por los efectos del encuestador y de la edad de manera flexible.

considerando que los cambios en el desarrollo cognitivo, de lenguaje y motor en los dos primeros años de vida se dan de manera muy rápida. De esta manera, hasta los 21 meses de edad se agrupan en un rango de 2 meses y, a partir de los 21 meses, el rango de edad se va incrementando en intervalos de hasta 5 meses. En esta desagregación, solo algunos coeficientes de la versión del ASQ-3 con 9 ítems tienen valores iguales o superiores al punto de corte (α=[0,70-0,84]), sobre todo en las escalas de comunicación y motricidad gruesa y no se concentran en ningún grupo etario en particular. Tres escalas tienen $\alpha \geq 0{,}80$: comunicación en LB (niños 22-25 meses), motricidad gruesa en LB (niños 13-15 meses) y comunicación en LS (niños 49-55 meses). Personal-social es la escala que, de manera global, presenta los valores de α más bajos. La consistencia interna de la versión con 9 ítems es para todas las escalas y todos los grupos etarios superior a la versión con 6 ítems. En parte, ello puede ser debido al mayor número de ítems. No se observa ningún patrón consistente respecto a la edad.

[Tabla 4]

Los resultados de consistencia interna para el Bayley-III son generalmente altos y por encima del punto de corte: α =[0,79-0,85] para desarrollo cognitivo; α =[0,77-0,81] para lenguaje receptivo; α =[0,81-0,89] para lenguaje expresivo y α =[0,81-0,88] para motricidad fina (Tabla 4). Al igual que el ASQ-3, la consistencia interna no presenta ningún patrón sistemático respecto a la edad.

5.2.2. Correlación entre las escalas

La Tabla 5 muestra las correlaciones de Pearson entre las escalas de una misma prueba.

[Tabla 5]

Se observan correlaciones bajas para todas las escalas del ASQ-3: entre r =[0,24-0,36] en la LB y r =[0,19-0,30] en la LS. Las correlaciones entre las escalas del Bayley-III son, por lo general, superiores a las correlaciones observadas entre las escalas del ASQ-3. Tres de ellas se encuentran en el rango moderado: desarrollo cognitivo con lenguaje receptivo (r =0,48) y con motricidad fina (r =0,44); y lenguaje receptivo con motricidad fina (r =0,41). Las otras correlaciones son bajas, entre r =[0,24-0,35].

5.3. Análisis de validez

5.3.1. Validez concurrente con variables socioeconómicas

La Tabla 6 reporta las correlaciones entre las escalas del ASQ-3 en LB y las del Bayley-III con variables socioeconómicas medidas en LB y que teóricamente están relacionadas con el desarrollo del niño.

[Tabla 6]

Por lo general, las correlaciones con el ASQ-3 tienen el signo esperado: el peso al nacer, la talla por edad, el peso por edad, la educación de la madre y del padre, el hecho que la madre trabaje, el índice de riqueza y los indicadores del FCI se asocian de manera positiva con el desarrollo del niño, mientras que para el tamaño del hogar se observan correlaciones negativas. Sin embargo, no todas estas correlaciones son significativas. Por ejemplo, las correlaciones entre el peso al nacer con las escalas de comunicación y personal-social; el hecho que la madre trabaje con los puntajes de resolución de

problemas, comunicación y motora fina; y el tamaño del hogar con las escalas de motora fina y gruesa no son estadísticamente significativas. En términos de magnitud, todos los coeficientes son muy bajos (todos con $r<0,20$ y la mayoría con $r<0,10$). Las correlaciones más altas son aquellas entre la educación de la madre con las escalas de resolución de problemas y personal-social ($r=0,15$ y $r=0,18$, respectivamente), la educación del padre con personal-social ($r=0,14$) y las actividades de juego con comunicación y personal-social ($r=0,19$ y $r=0,17$ respectivamente). La edad de la madre no se correlaciona con ninguna de las escalas analizadas.

Se observa un patrón de resultados similar para el Bayley-III, si bien existen algunas diferencias. Por ejemplo, entre las correlaciones significativas, solo la educación de la madre y del padre se correlacionan con *todas* las escalas de la prueba. A diferencia del ASQ-3, la talla por edad y las dos escalas del FCI solo se correlacionan con desarrollo cognitivo y lenguaje receptivo, y el peso por edad no se correlaciona con lenguaje expresivo. Las magnitudes de las correlaciones para el Bayley-III son mucho más altas que las del ASQ-3 con las mismas variables. Específicamente, esto se observa para la talla y el peso por edad, la educación de la madre y del padre, y el índice de riqueza.

5.3.2. Validez concurrente entre las escalas de las dos pruebas

En la Tabla 7 se exhiben las correlaciones entre las escalas del ASQ-3 en la LS y las del Bayley-III (27-43 meses de edad para el ASQ-3 y 26-42 meses para el Bayley-III). Aquellas correlaciones entre las escalas que miden el mismo dominio del desarrollo están resaltadas en negritas.

[Tabla 7]

Se observan solo cuatro correlaciones bajas: resolución de problemas, comunicación y motricidad fina del ASQ-3 con lenguaje expresivo del Bayley-III ($r=0,24$; $r=0,38$; y $r=0,26$, respectivamente) y motricidad fina del ASQ-3 con motricidad fina del Bayley-III ($r=0,28$). De estas cuatro, dos miden los mismos dominios del desarrollo: comunicación con lenguaje expresivo ($r=0,38$) y las dos escalas de motricidad fina ($r=0,28$). La escala de comunicación del ASQ-3 tiene una correlación de más del doble con lenguaje expresivo ($r=0,38$) comparado con lenguaje receptivo ($r=0,15$); si bien lenguaje expresivo es la escala de Bayley-III que presenta las correlaciones más bajas con las variables socio-económicas (Tabla 6). Resolución de problemas tiene una correlación más alta con lenguaje expresivo ($r=0,24$) que con la escala de desarrollo cognitivo ($r=0,15$), a pesar de que miden los mismos constructos. La escala personal-social del ASQ-3 no se correlaciona con ninguna de las escalas del Bayley-III que fueron administradas.

5.3.3 Validez predictiva

La Tabla 8 presenta los resultados de las correlaciones entre las escalas del ASQ-3 en LB y las escalas del ASQ-3 en LS (panel de la izquierda), y del ASQ-3 en LB con las escalas del Bayley-III (panel de la derecha; 27-43 meses de edad para el ASQ-3, y 26-42 para el Bayley-III).

[Tabla 8]

Los resultados muestran correlaciones y una capacidad predictiva muy bajas. La escala de resolución de problemas en LB tiene una correlación más alta con la escala de comunicación en la LS ($r=0,18$) y con lenguaje receptivo y lenguaje expresivo del

Bayley-III (r=0,24 y r=0,20, respectivamente) que con su misma escala en la LS (r=0,15). La validez predictiva de la escala de comunicación muestra los valores más altos con comunicación en la LS (r=0,17) y con lenguaje expresivo (r=0,20) del Bayley-III, pero no se correlaciona de manera significativa con lenguaje receptivo. La escala de motricidad fina tiene la correlación más alta con motricidad fina del ASQ-3 en LS de todas las escalas de las dos pruebas (r=0,16), pero no tiene una correlación significativa con motricidad fina del Bayley-III. La escala de personal-social en LB no se correlaciona con su misma escala en la LS, ni tampoco con motricidad gruesa. Finalmente, la escala de motricidad gruesa en LB tiene la correlación más alta con lenguaje receptivo (r=0,17) y no se correlaciona con personal-social en la LS.

6. Discusión

Este estudio describe el proceso de adaptación de una prueba de tamizaje, el ASQ-3, y de una prueba de diagnóstico, el Bayley-III, ambas administradas en distritos rurales con altos niveles de pobreza en el Perú, y analiza sus propiedades psicométricas. Es importante tener en cuenta que los datos no fueron recolectados con el propósito de hacer un ejercicio de validación de las pruebas, sino en el contexto de la evaluación de impacto del SAF. A pesar de que las pruebas de diagnóstico se consideran como las más adecuadas para la medición del DIT, al diseñar la evaluación se optó por usar el ASQ-3 porque representaba una alternativa menos costosa y más fácil de administrar, había sido usada con ese mismo objetivo en Perú, y existía una versión en español publicada. En la medición de seguimiento, se decidió administrar adicionalmente el Bayley-III a una submuestra de niños para contar con una medición más fina del DIT. Ambas pruebas pasaron por un proceso de adaptación riguroso al contexto rural peruano y a poblaciones vulnerables. El ASQ-3 fue administrado por encuestadores en el hogar, combinando la administración directa con el reporte materno (vía entrevista). El Bayley-III, en cambio, fue administrado por psicólogas a través de la observación del desempeño del niño en la realización de una serie de ítems, en un lugar acondicionado para ello en la comunidad.

La comparación de nuestros resultados con otros estudios se ve limitada dado que ninguna de las experiencias previas del ASQ y del Bayley en Perú revisadas reportaron resultados de confiabilidad o de validez de las pruebas (a excepción del alfa de Cronbach en Fernald et al. 2012). En la medida de lo posible, comparamos nuestros hallazgos con aquellos reportados en los manuales de las pruebas. También hacemos una comparación con los resultados del estudio reciente de Rubio-Codina et al. (2016a), que hicieron un ejercicio similar en Bogotá.

La consistencia interna del ASQ-3 mostró por lo general valores inferiores al punto de corte deseado (α<0,70) para la LB y la LS, tanto para la versión de 6 como para la de 9 ítems. Sin embargo, los resultados no difieren mucho de aquellos reportados en el manual de la prueba y en otros estudios en contextos similares, sobre todo para la versión con 9 ítems en la LB. Por ejemplo, para la LB (versión con 9 ítems), las escalas de resolución de problemas, motricidad fina y personal-social tienen α iguales o superiores a las del manual, mientras que las de comunicación y motricidad gruesa son levemente inferiores (α =0,67 vs. α =0,71; y α =0,68 vs. α =0,72, respectivamente).[10] Fernald et al. (2012) y Rubio-Codina et al. (2016a) reportan valores parecidos. En cambio, para el

[10] La Tabla 15 del manual del ASQ-3 reporta el alfa por cuestionario para cada escala. Para que los resultados sean comparables, calculamos el promedio de las alfas de los mismos cuestionarios que se usaron en la LB y en la LS.

Bayley-III, la consistencia interna fue aceptable en tanto todas las escalas tienen un $\alpha > 0,8$.

Las correlaciones entre las escalas del ASQ-3 se encontraron en el rango bajo (r=[0,24-0,36] en la LB; y r=[0,19-0,30] en la LS). Estos valores son más bajos que aquellos reportados en el manual (r=[0,33-0,54]). Esta comparación es limitada dado que las correlaciones en el manual abarcan edades entre 1 a 66 meses.

Para el Bayley-III, las correlaciones fueron algo mayores, entre los rangos bajo y moderado (r=[0,24-0,48]). Los valores para las escalas de desarrollo cognitivo, lenguaje receptivo y motricidad fina se acercan a los valores del manual. En cambio, los valores para la escala de lenguaje expresivo son mucho menores (r=0,24 vs. r=0,45 con desarrollo cognitivo; y r=0,35 vs. r=0,53 con lenguaje receptivo). Rubio-Codina et al. (2016a) en Bogotá encontraron valores parecidos a los del manual. Las condiciones de administración del Bayley-III fueron muy similares en el estudio de Perú y en el de Bogotá. En ambos casos, se seleccionó un número limitado de psicólogas (8 y 6, respectivamente) con experiencia en DIT, que fueron capacitadas por 6 semanas, y que administraron la prueba en condiciones óptimas, cumpliendo con los requisitos de administración de la prueba original. En el caso de Perú, los índices de confiabilidad intersujeto durante el operativo de campo fueron incluso más altos que los del piloto, entre 96-99% de acuerdos. En cambio, la población de estudio fue muy diferente: mientras que la muestra de Bogotá se concentró en una población urbana, la de Perú se enfocó en una población rural. A pesar de que se hizo un ejercicio muy riguroso de adaptación de la prueba en el estudio de Perú, es posible que el instrumento tenga más limitaciones cuando se administra en poblaciones tan diversas.

En el análisis de validez de criterio, las correlaciones entre el ASQ-3 y variables socioeconómicas fueron muy bajas (todas r<0,20 y la mayoría r<0,10). Se observó un patrón similar para el Bayley-III, pero las magnitudes de las correlaciones fueron mayores, en algunos casos el doble que las del ASQ-3 (aun así, en el rango bajo). Se destaca que las correlaciones con los materiales y las actividades de juego del FCI son mucho menores a los valores que se encontraron en el estudio de Bogotá para ambas pruebas. A modo de ejemplo, las correlaciones entre los materiales de juego y desarrollo cognitivo, lenguaje receptivo, lenguaje expresivo y motricidad fina del Bayley-III se encuentran entre r=[0,18-0,27] en el estudio de Bogotá, mientras que en Perú solo dos de ellas son significativas (desarrollo cognitivo y lenguaje receptivo) y con una magnitud de r=0,11. Si bien los estudios de Perú y de Bogotá fueron muy similares en algunos aspectos (como la capacitación de los encuestadores que administraron el ASQ-3, sus perfiles y las condiciones de administración de la encuesta de hogar) también hubo diferencias. Por ejemplo, el número de encuestadores era 8 veces mayor en la LB del estudio de Perú, donde se capacitó a un total de 76 personas quienes administraron el ASQ-3 y el FCI. Es posible que esto influyó en la calidad de los datos del FCI. Además, la disponibilidad de materiales de juego y el número de actividades realizadas eran mucho menor en la muestra de Perú en comparación con la muestra de Bogotá. Es posible que, como resultado de la menor variabilidad en la muestra de Perú, estas variables tengan un menor poder explicativo y que esto se refleje en menores correlaciones con los puntajes de las pruebas de DIT.

La capacitación de los encuestadores es un elemento clave para preservar la calidad de los datos, especialmente en encuestas de esta escala. Es importante tomar todo el tiempo necesario para los procesos de capacitación, aunque ello incremente los costos. El trabajar con grupos pequeños, incluir pilotos en campo y evaluar la confiabilidad intersujeto son buenas prácticas para el aseguramiento de la calidad de los datos. En algunas pruebas como el Bayley, podría incluso ser necesario analizar dicha confiabilidad al nivel de las escalas, no solo con un puntaje global.

En cuanto a la validez predictiva, las correlaciones entre las escalas coincidentes del ASQ-3 en LB y LS se encontraron todas en el rango muy bajo (r=[0,12-0,16]) y, en el caso de la escala de personal-social, no fueron significativas (r=0,06). En cuanto a las correlaciones entre el ASQ-3 en LB y el Bayley-III, solo dos de ellas fueron significativas: resolución de problemas y desarrollo cognitivo y comunicación y lenguaje expresivo—ambas muy bajas. Si bien la validez predictiva y la validez concurrente no siempre están estrechamente relacionadas, este resultado no es tan sorprende considerando los niveles de validez concurrente bajos que mostró la prueba en el presente estudio y también en el de Bogotá en niños menores de 31 meses (Rubio-Codina et al. 2016a). Esto puede ser una indicación de que la prueba, a pesar de tener muchas ventajas en cuanto a la facilidad de capacitación de los encuestadores y de su administración, al ser una prueba de tamizaje, tiene una capacitad limitada para predecir el desarrollo futuro de niños muy pequeños.

El Bayley-III, al ser una prueba más compleja y costosa, de diagnóstico, que busca establecer el máximo rendimiento posible del niño en determinada edad, tiene en general mejores propiedades. Es necesario seguir investigando en el país y la región, con diversos instrumentos, para encontrar alternativas viables y confiables para medir los impactos de programas de DIT que operan a escala.

Referencias

Ángeles, G., Gadsden, P., Galiani, S., Gertler, P., Herrera, A., Kariger, P. y Seira, E., 2011. Evaluación de impacto del programa estancias infantiles para apoyar a madres trabajadoras. *México, DF*

Attanasio, O.P., Fernández, C., Fitzsimons, E.O., Grantham-McGregor, S.M., Meghir, C. y Rubio-Codina, M., 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*, p.g5785

Araujo, M.C., Dormal, M. Grantham-McGregor, S., Lazarte, F., Rubio-Codina, M., y Schady, N. 2018. Behavioral change and child development at scale. Manuscrito no publicado.

Bernal, R., 2015. The impact of a vocational education program for childcare providers on children's well-being. *Economics of Education Review*, *48*, pp.165-183.

Banco Interamericano de Desarrollo. 2017. Estrategia del grupo BID con Perú (2017-2021). Washington, D.C.: Banco Interamericano de desarrollo.

Banco Mundial. 2017a. "World Development Indicators." Disponible en: http://databank.bancomundial.org/data/source/world-development-indicators Consultado el 16 de marzo 2017.

Banco Mundial. 2017b. Peru - Systematic Country Diagnostic (English). Washington, D.C.: World Bank Group.

Bayley, N., 2006. *Bayley Scales of Infant and Toddler Development: Technical manual* (3ra ed.). San Antonio, TX: Harcourt Assessment.

Berlinski, S. y Schady, N. (eds.) 2015. *The Early Years: Child Well-being and the Role of Public Policy. Development in the Americas Series*. New York: Palgrave MacMillan y Washington, DC: Inter-American Development Bank.

Blouin, B., Casapia, M., Joseph, L. y Gyorkos, T.W., 2018. A longitudinal cohort study of soil-transmitted helminth infections during the second year of life and associations with reduced long-term cognitive and verbal abilities. *PLoS neglected tropical diseases*, *12*(7), p.e0006688.

Bricker, D., Squires, J. y Mounts, L., 1999. *Ages and Stages Questionnaires (ASQ): A parent-completed, child-monitoring system* (2nd ed.) Baltimore: Paul H. Brookes Publishing Co.

Colombo, J., Zavaleta, N., Kannass, K.N., Lazarte, F., Albornoz, C., Kapa, L.L. y Caulfield, L.E. 2014. Zinc Supplementation Sustained Normative Neurodevelopment in a Randomized, Controlled Trial of Peruvian Infants Aged 6–18 Months, 2. *The Journal of nutrition*, *144*(8), pp.1298-1305.

Cunha, F., Heckman, J.J., Lochner, L. y Masterov, D.V., 2006. Interpreting the evidence on life cycle skill formation. *Handbook of the Economics of Education*, *1*, pp.697-812.

Currie, J., 2009. Healthy, wealthy, and wise: Socioeconomic status, poor health in childhood, and human capital development. *Journal of economic literature*, *47*(1), pp.87-122.

DGSE-MIDIS, DGPP-MEF y BID, 2015. Documento de Línea de Base para la Evaluación de Impacto del Programa Nacional Cuna Más – Servicio de Acompañamiento a Familias. Documento interno del MIDIS.

Evans, J. D., 1996. Straightforward statistics for the behavioral sciences. Pacific Grove: Brooks/Cole Pub

Fernald, L.C. y Hidrobo, M., 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 & medicine*, *72*(9), pp.1437-1446.

Fernald, L.C., Kariger, P., Hidrobo, M. y Gertler, P.J., 2012. Socioeconomic gradients in child development in very young children: Evidence from India, Indonesia, Peru, and Senegal. *Proceedings of the National Academy of Sciences*, p.201121241.

Fernald, L.C., Prado, E., Kariger, P., y Raikes, A, 2017. *A toolkit for measuring early childhood development in low and middle income countries*. Washington, D.C.: World Bank Group.

Fernandes, M., Stein, A., Newton, C.R., Cheikh-Ismail, L., Kihara, M., Wulff, K., de León Quintana, E., Aranzeta, L., Soria-Frisch, A., Acedo, J. y Ibanez, D., 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), p.e113360.

Frongillo, E.A., Tofail, F., Hamadani, J.D., Warren, A.M. y Mehrin, S.F., 2014. Measures and indicators for assessing impact of interventions integrating nutrition, health, and early childhood development. *Annals of the New York academy of sciences*, *1308*(1), pp.68-88

Gertler, P., Heckman, J., Pinto, R., Zanolini, A., Vermeersch, C., Walker, S., Chang, S.M. y Grantham-McGregor, S., 2014. Labor market returns to an early childhood stimulation intervention in Jamaica. *Science*, *344*(6187), pp.998-1001.

Grantham-McGregor, S., Cheung, Y.B., Cueto, S., Glewwe, P., Richter, L., Strupp, B. y International Child Development Steering Group, 2007. Developmental potential in the first 5 years for children in developing countries. *The Lancet*, *369*(9555), pp.60-70.

Grantham-McGregor, S.M., Powell, C.A., Walker, S.P. y Himes, J.H., 1991. Nutritional supplementation, psychosocial stimulation, and mental development of stunted children: the Jamaican Study. *The Lancet*, *338*(8758), pp.1-5.

Hamadani, J.D., Baker-Henningham, H., Tofail, F., Mehrin, F., Huda, S.N. y Grantham-McGregor, S.M., 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_suppl2), pp.S198-S206.

Hamadani, J.D., Tofail, F., Cole, T. y Grantham-McGregor, S., 2013. The relation between age of attainment of motor milestones and future cognitive and motor development in B angladeshi children. *Maternal & child nutrition*, *9*, pp.89-104.

Heckman, J.J., Stixrud, J. y Urzua, S., 2006. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor economics*, *24*(3), pp.411-482.

Heckman, J.J., 2007. The economics, technology, and neuroscience of human capability formation. *Proceedings of the national Academy of Sciences*, *104*(33), pp.13250-13255

INEI (Instituto Nacional de Estadística e Informática), 2017. Estadísticas Sociales. Disponible en http://www.inei.gob.pe/estadisticas/indice-tematico/sociales/. Consultado el 16 de marzo 2017.

Joseph, S.A., Casapía, M., Lazarte, F., Rahme, E., Pezo, L., Blouin, B. y Gyorkos, T.W., 2015. The effect of deworming on early childhood development in Peru: A randomized controlled trial. *SSM-population health*, *1*, pp.32-39.

Kariger, P., Frongillo, E.A., Engle, P., Britto, P.M.R., Sywulka, S.M. y Menon, P. 2012. "Indicators of family care for development for use in multicountry surveys". *Journal of health, population, and nutrition*, 30(4), p.472.

Kyerematen, V., Hamb, A., Oberhelman, R.A., Cabrera, L., Bernabe-Ortiz, A. y Berry, S.J., 2014. Exploratory application of the Ages and Stages (ASQ) child development screening test in a low-income Peruvian shantytown population. *BMJ open*, *4*(1), p.e004132.

Maluccio, J.A., Hoddinott, J., Behrman, J.R., Martorell, R., Quisumbing, A.R. y Stein, A.D., 2009. The impact of improving nutrition during early childhood on education among Guatemalan adults. *The Economic Journal*, *119*(537), pp.734-763.

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

MAL-ED Network Investigators, 2014. The MAL-ED study: A multinational and multidisciplinary approach to understand the relationship between enteric pathogens, malnutrition, gut physiology, physical growth, cognitive development, and immune responses in infants and children up to 2 years of age in resource-poor environments". *Clinical Infectious Diseases*, 59, S193–S206.

Nahar, B., Hamadani, J.D., Ahmed, T., Tofail, F., Rahman, A., Huda, S.N. y Grantham-McGregor, S.M., 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), p.725.

Nelson, A.K., Miller, A.C., Munoz, M., Rumaldo, N., Kammerer, B., Vibbert, M., Lundy, S., Soplapuco, G., Lecca, L., Condeso, A. y Valdivia, Y., 2018. CASITA: a controlled pilot

study of community-based family coaching to stimulate early child development in Lima, Peru. *BMJ paediatrics open*, *2*(1).

Ore, B., Díaz, J.J. y Penny, M., 2011. Impacto de una intervención con grupos de mamás y bebes en el desarrollo infantil. *Revista de Psicología (PUCP)*, *29*(1), pp.37-66.

Pollit, E. y Granoff, D., 2017. Mental and motor development of Peruvian children treated for severe malnutrition. *Revista Interamericana de Psicologia/Interamerican Journal of Psychology*, *1*(2).

Rubio-Codina, M., Araujo, M.C., Attanasio, O., Muñoz, P. y Grantham-McGregor, S., 2016. Concurrent validity and feasibility of short tests currently used to measure early childhood development in large scale studies. *PloS one*, *11*(8), p.e0160962.

Rubio-Codina, M., Attanasio, O. y Grantham-McGregor, S., 2016. Mediating pathways in the socio-economic gradient of child development: Evidence from children 6–42 months in Bogota. *International journal of behavioral development*, *40*(6), pp.483-491

Scharf, R.J., Rogawski, E.T., Murray-Kolb, L.E., Maphula, A., Svensen, E., Tofail, F., Rasheed, M., Abreu, C., Vasquez, A.O., Shrestha, R. y Pendergast, L., 2018. Early childhood growth and cognitive outcomes: Findings from the MAL-ED study. *Maternal & child nutrition*, p.e12584.

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

Schady, N., Behrman, J.. Araujo, M.C., Azuero, R., Bernal, R., Bravo, D., Lopez-Boo, F., Macours, K., Marshall, D., Paxson, C. y Vakis, R., 2015. Wealth gradients in early childhood cognitive development in five Latin American Countries. *Journal of Human Resources*, 50(2): 446-463.

Walker, S.P., Chang, S.M., Vera-Hernández, M. y Grantham-McGregor, S., 2011. Early childhood stimulation benefits adult competence and reduces violent behavior. *Pediatrics*, pp.peds-2010.

Westgard, C. y Alnasser, Y., 2017. Developmental delay in the Amazon: The social determinants and prevalence among rural communities in Peru. *PloS one*, *12*(10), p.e0186263.

WHO Multicentre Growth Reference Study Group, 2006. WHO Child Growth Standards: Length/height-for-age, weight-for-age, weight-for-length, weight-for-height and body mass index-for-age: Methods and development. Geneva: World Health Organization"; pp 312. Disponible en http://www.who.int/childgrowth/publication

Tabla 1: Características de los niños en las muestras ASQ-3 y Bayley-III

	ASQ-3 (N=1.493)			Bayley-III (N=385)			
	N	M/ Proporción	DE	N	M/ Proporción	DE	P-valor
Características de los niños							
Edad (meses)	1.493	12,81	6,59	385	7,54	3,68	0,00
Mujeres (%)	1.493	0,49	0,50	385	0,54	0,50	0,02
Peso al nacer (gr)	820	3144,63	433,20	243	3137,13	443,14	0,75
Prematuro (edad gestacional <37 semanas, %)	1.493	0,07	0,25	385	0,08	0,27	0,30
Desnutrición crónica (puntaje z de talla por edad <-2DE, %)	1.440	0,36	0,48	371	0,31	0,46	0,01
Desnutrición global (puntaje z de peso por edad <-2DE, %)	1.441	0,07	0,25	371	0,05	0,23	0,21
Primogénito (%)	1.493	0,29	0,46	385	0,30	0,46	0,94
Niño recibió leche materna durante 6 meses o más (%)	1.477	0,73	0,44	380	0,73	0,44	0,90
Características de las familias							
Edad de la madre (años)	1.493	28,37	7,61	385	27,81	6,97	0,10
Educación de la madre (años)	1.417	6,69	4,02	369	6,93	4,22	0,18
Madre trabaja (%)*	1.493	0,37	0,48	385	0,36	0,48	0,47
Lengua materna de la madre es indigena (%)	1.493	0,31	0,46	385	0,32	0,47	0,43
Educación del padre (años)**	1.227	7,88	3,87	316	7,81	3,96	0,74
Características del hogar							
Activos y aceso a servicios (%)							
Equipo de sonido	1.493	0,14	0,35	385	0,14	0,35	0,79
Televisor	1.493	0,47	0,50	385	0,51	0,50	0,06
DVD player	1.493	0,33	0,47	385	0,36	0,48	0,19
Licuadora	1.493	0,21	0,41	385	0,18	0,39	0,17
Refrigeradora	1.493	0,08	0,27	385	0,06	0,25	0,25
Cocina a gas	1.493	0,40	0,49	385	0,41	0,49	0,52
Tabla de planchar	1.493	0,13	0,33	385	0,12	0,32	0,45
Celular	1.493	0,63	0,48	385	0,70	0,46	0,00
Piso distinto de tierra	1.493	0,28	0,45	385	0,27	0,44	0,51
Electricidad	1.493	0,78	0,41	385	0,79	0,41	0,61
Agua (dentro de la vivienda)	1.493	0,59	0,49	385	0,57	0,50	0,37
Servicio higiénico (dentro de la vivienda)	1.493	0,22	0,42	385	0,25	0,43	0,19
Hacinamiento	1.493	2,81	1,58	385	2,80	1,61	0,94
Tamaño del hogar	1.493	5,57	2,09	385	5,49	2,02	0,37
Ubicación geográfica (%)***							
Costa urbana	-	-	-	-	-	-	-
Costa rural	-	-	-	-	-	-	-
Sierra urbana	1.493	0,20	0,40	385	0,26	0,44	0,00
Sierra rural	1.493	0,57	0,50	385	0,57	0,50	0,94
Selva urbana	1.493	0,11	0,31	385	0,00	0,00	0,00
Selva rural	1.493	0,12	0,33	385	0,17	0,37	0,00
Calidad del ambiente del hogar							
Número de materiales de juego (puntaje crudo, 0-8)	1.458	1,67	1,33	376	1,31	1,20	0,00
Componentes de los materiales de juego (%):							
Juguetes con los que se producen o se toca música	1.298	0,11	0,32	310	0,13	0,33	0,42
Juguetes para armar o construir	1.299	0,14	0,34	310	0,07	0,26	0,00
Cosas para pintar o escribir	1.299	0,09	0,28	310	0,03	0,16	0,00
Juguetes que requieren mucho movimiento físico	1.299	0,66	0,47	310	0,53	0,50	0,00
Muñecos, muñecas y objetos para el juego de roles o juego de fantasías	1.299	0,67	0,47	310	0,67	0,47	0,96
Libros infantiles para colorear	1.299	0,04	0,20	310	0,02	0,13	0,01
Libros infantiles de cuentos	1.297	0,06	0,24	309	0,06	0,23	0,77
Juguetes para aprender formas y/o colores	1.298	0,10	0,30	309	0,08	0,28	0,26
Número de actividades de juego (puntaje crudo, 0-7)	1.493	2,86	1,86	385	2,46	1,68	0,00
Componentes de las actividades de juego (%):							
Leer libros, ver cuentos de imágenes o dibujos de un libro	1.493	0,27	0,45	385	0,20	0,40	0,00
Contarle cuentos o historias	1.493	0,19	0,40	385	0,18	0,38	0,31
Cantar canciones	1.493	0,56	0,50	385	0,51	0,50	0,04
Salir a pasear	1.493	0,64	0,48	385	0,59	0,49	0,03
Jugar con sus juguetes	1.493	0,70	0,46	385	0,70	0,46	0,82
Dibujar, pintar, escribir o jugar a hacer garabatos en papel	1.493	0,27	0,44	385	0,14	0,35	0,00
Jugar a nombrar objetos o colores, a contar objetos o a decir los números	1.493	0,23	0,42	385	0,14	0,35	0,00
Hogares con libros para adultos (%)	1.493	0,32	0,47	382	0,35	0,48	0,17

Nota: *Se preguntó si la madre había trabajado en la semana anterior a la encuesta
** Solo para los padres que vivían en el hogar en el momento de la encuesta
*** Según la definición del Instituto Nacional de Estadísticas e Informática del Perú

Tabla 2: Puntajes brutos del ASQ-3 y puntajes crudos y compuestos del Bayley-III

	ASQ-3 (N=1.493)			**Bayley-III** (N=385)			
	N	Media	DE	N	Media	DE	P-valor
ASQ-3 (9 ítems, línea de base)							
Resolución de problemas	1.493	48,58	18,01	385	49,01	18,91	0,59
Comunicación	1.493	43,93	18,57	385	49,92	17,16	0,00
Motora fina	1.493	47,45	17,66	385	50,70	18,48	0,00
Personal-social	1.493	47,49	16,37	385	47,31	15,86	0,80
Motora gruesa	1.493	43,98	18,40	385	41,30	17,38	0,00
ASQ-3 (9 ítems, línea de seguimiento)							
Resolución de problemas	1.493	34,61	13,82	385	36,81	13,04	0,00
Comunicación	1.493	41,24	15,80	385	41,17	13,62	0,92
Motora fina	1.493	39,83	16,00	385	42,22	14,16	0,00
Personal-social	1.493	52,40	18,33	385	51,97	16,99	0,59
Motora gruesa	1.493	46,81	16,90	385	46,00	17,16	0,28
Bayley-III							
Puntaje crudos							
Desarrollo cognitivo	-	-	-	385	67,41	4,52	-
Lenguaje receptivo	-	-	-	385	31,03	3,65	-
Lenguaje expresivo	-	-	-	385	32,71	3,99	-
Motora fina	-	-	-	385	45,10	5,05	-
Puntaje compuestos							
Cognitivo	-	-	-	385	86,97	6,36	-
Lenguaje	-	-	-	385	88,90	5,85	-

Nota: los puntajes compuestos corresponden a una función no lineal de los puntajes brutos, con una media de 100 y una desviación estándar de 15. Las escalas de lenguaje receptivo y expresivo se combinan para generar un puntaje único sobre lenguaje.

Tabla 3: Consistencia interna (alfa de Cronbach) para el ASQ-3, muestra completa y por grupo etario

	Resolución de problemas		Comunicación		Motora fina		Personal-social		Motora gruesa	
	6 ítems	9 ítems	6 ítems	9 ítems	6 ítems	9 ítems	6 ítems	9 ítems	6 ítems	9 ítems
Línea de base										
muestra completa (n=1.493)	0,38	0,66	0,42	0,67	0,39	0,63	0,31	0,60	0,55	0,68
1-3 meses (n=135)	0,63	0,70	0,43	0,75	0,26	0,57	0,32	0,59	0,57	0,71
4-6 meses (n=193)	0,52	0,74	0,29	0,64	0,42	0,76	0,35	0,57	0,41	0,52
7-9 meses (n=199)	0,36	0,66	0,33	0,62	0,49	0,72	0,34	0,64	0,63	0,72
10-12 meses (n=197)	0,38	0,66	0,17	0,63	0,39	0,58	0,37	0,58	0,60	0,67
13-15 meses (n=182)	0,38	0,71	0,47	0,60	0,52	0,62	0,29	0,66	0,66	0,81
16-18 meses (n=214)	0,44	0,70	0,44	0,59	0,44	0,61	0,31	0,73	0,69	0,79
19-21 meses (n=192)	0,22	0,57	0,59	0,71	0,24	0,54	0,20	0,44	0,52	0,65
22-25 meses (n=181)	0,19	0,57	0,69	0,84	0,34	0,59	0,28	0,54	0,31	0,58
Línea de seguimiento										
muestra completa (n=1.493)	0,41	0,47	0,46	0,62	0,40	0,55	0,26	0,70	0,35	0,63
25-30 meses (n=95)	0,61	0,71	0,40	0,74	0,29	0,60	0,32	0,59	0,55	0,70
31-36 meses (n=373)	0,47	0,70	0,33	0,65	0,42	0,72	0,34	0,60	0,52	0,62
37-42 meses (n=394)	0,38	0,67	0,31	0,61	0,45	0,62	0,33	0,62	0,63	0,73
43-48 meses (n=407)	0,35	0,65	0,50	0,64	0,37	0,59	0,27	0,61	0,61	0,73
49-55 meses (n=224)	0,20	0,57	0,66	0,80	0,32	0,58	0,26	0,52	0,37	0,60

Nota: Dado que el ASQ-3 tiene cuestionarios que son edad-específicos, calculamos primero las alfas de Cronbach por cuestionario (i.e., se le asigna un alfa a cada niño según el cuestionario que le fue administrado), y después calculamos el promedio de las alfas para los niños en cada grupo etario.

Tabla 4: Consistencia interna (alfa de Cronbach) para el Bayley-III, muestra completa y por grupo etario

	Desarrollo cognitivo	Lenguaje receptivo	Lenguaje expresivo	Motora fina
muestra completa (n=385)	0,85	0,81	0,86	0,88
26-30 meses (n=41)	0,81	0,80	0,89	0,83
31-36 meses (n=186)	0,79	0,77	0,85	0,81
37-42 meses (n=158)	0,84	0,77	0,81	0,88

Tabla 5: Correlaciones entre las escalas de una misma prueba

ASQ-3 (n=1.493)	**Línea de base**				
	Resolución de Problemas	Comunicación	Motora Fina	Personal-social	Motora Gruesa
Línea de base					
Resolución de problemas	1				
Comunicación	0,24***	1			
Motora fina	0,36***	0,25***	1		
Personal-social	0,28***	0,30***	0,28***	1	
Motora gruesa	0,25***	0,25***	0,28***	0,28***	1
	Línea de seguimiento				
Línea de seguimiento					
Resolución de problemas	1				
Comunicación	0,30***	1			
Motora fina	0,28***	0,25***	1		
Personal-social	0,21***	0,26***	0,21***	1	
Motora gruesa	0,23***	0,22***	0,21***	0,19***	1

Bayley-III (n=385)	**Bayley-III**			
	Desarrollo cognitivo	Lenguaje receptivo	Lenguaje Expresivo	Motora fina
Desarrollo cognitivo	1			
Lenguaje receptivo	0,48***	1		
Lenguaje expresivo	0,24***	0,35***	1	
Motora fina	0,44***	0,41***	0,32***	1

Nota: coeficientes de correlación de Pearson. Significativos al *p<0,10, ** p<0,05 y *** p<0,01.

Tabla 6: Correlaciones entre el ASQ-3 y el Bayley-III con variables socioeconómicas

	ASQ-3 en línea de base (n=1.493)					Bayley-III (n=385)			
Características de los niños	Resolución de problemas	Comunicación	Motora fina	Personal-social	Motora gruesa	Desarrollo cognitivo	Lenguaje receptivo	Lenguaje expresivo	Motora fina
Peso al nacer (gr)	0,10***	0,02	0,11***	0,03	0,07**	0,11*	0,03	0,10	0,05
Talla por edad (puntaje z)	0,09***	0,12***	0,12***	0,10***	0,12***	0,20***	0,16***	0,07	0,05
Peso por edad (puntaje z)	0,08***	0,10***	0,09***	0,08***	0,11***	0,20***	0,21***	0,06	0,13**
Niño recibió leche materna durante 6 meses o más	-0,01	0,00	0,00	-0,02	-0,02	0,01	-0,03	0,02	-0,05
Características de las familias									
Edad de la madre (años)	-0,03	-0,03	0,00	-0,03	-0,03	0,03	0,01	0,01	0,02
Educación de la madre (años)	0,15***	0,11***	0,11***	0,18***	0,10***	0,27***	0,23***	0,17***	0,26***
Madre trabaja	0,02	0,04	0,01	0,07***	0,08***	0,07	0,03	0,07	0,09*
Educación del padre (años)	0,08***	0,09***	0,10***	0,14***	0,10***	0,25***	0,21***	0,17***	0,22***
Características del hogar									
Indice de riqueza	0,11***	0,06**	0,12***	0,14***	0,08***	0,19***	0,19***	0,07	0,22***
Tamaño del hogar	-0,07**	-0,07***	-0,04	-0,07***	-0,02	-0,07	-0,08	-0,08	0,01
Número de materiales de juego (puntaje z)	0,09***	0,10***	0,06**	0,09***	0,02	0,11**	0,11*	0,03	-0,02
Número de actividades de juego (puntaje z)	0,10***	0,19***	0,10***	0,17***	0,11***	0,09*	0,10*	0,05	0,08

Nota: coeficientes de correlación de Pearson. Significativos al *p<0,10, ** p<0,05 y *** p<0,01.

Tabla 7: Validez concurrente (N=385)

	Bayley-III			
	Desarrollo cognitivo	Lenguaje receptivo	Lenguaje Expresivo	Motora fina
ASQ-Línea de seguimiento				
Resolución de problemas	**0,15*** **	0,11**	0,24***	0,09*
Comunicación	0,11**	**0,15*** **	**0,38*** **	0,10**
Motora fina	0,17***	0,18***	0,26***	**0,28*** **
Personal-social	0,02	0,03	0,04	0,03
Motora gruesa	0,13***	0,09*	0,04	0,12**

Nota: coeficientes de correlación de Pearson. Significativos al *p<0,10, ** p<0,05 y *** p<0,01.

Tabla 8: Validez predictiva (N=385)

	ASQ-3-Línea de seguimiento					Bayley-III			
ASQ-3-Línea de base	Resolución de problemas	Comunicación	Motora fina	Personal-social	Motora gruesa	Desarrollo cognitivo	Lenguaje receptivo	Lenguaje Expresivo	Motora fina
Resolución de problemas	**0,15*** **	0,18***	0,14***	0,04	0,13***	**0,13** **	0,24***	0,20***	0,10*
Comunicación	0,11**	**0,17*** **	0,11**	0,11**	0,14***	0,09*	**0,08**	**0,20*** **	0,10**
Motora fina	0,05	0,12**	**0,16*** **	0,05	0,09**	0,13**	0,09*	0,04	**0,07**
Personal-social	0,14**	0,17***	0,15***	**0,06**	0,05	0,16**	0,15**	0,12**	0,14***
Motora gruesa	0,10**	0,10*	0,12**	-0,01	**0,12** **	0,14**	0,17***	0,10*	0,11**

Nota: coeficientes de correlación de Pearson. Significativos al *p<0,10, ** p<0,05 y *** p<0,01.

Apéndice A

Tabla A1: Porcentaje de niños que lograron la puntuación máxima en el ASQ-3 (N=1.493)

	6 ítems (puntaje = 60)		9 ítems (puntaje = 90)	
	N	%	N	%
Línea de base				
Resolución de problemas	302	20,23	70	4,69
Comunicación	219	14,67	71	4,76
Motora fina	279	18,69	72	4,82
Personal-social	236	15,81	63	4,22
Motora gruesa	237	15,87	46	3,08
Línea de seguimiento				
Resolución de problemas	63	4,22	4	0,27
Comunicación	122	8,17	22	1,47
Motora fina	138	9,24	18	1,21
Personal-social	345	23,11	148	9,91
Motora gruesa	248	16,61	65	4,35