IDB WORKING PAPER SERIES Nº IDB-WP-723

Concurrent Validity and Feasibility of Short Tests Currently Used to Measure Early Childhood Development in Large Scale Studies: Methodology and Results

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

Inter-American Development Bank
Social Protection and Health Division

August 2016

IDB
Inter-American
Development Bank

Concurrent Validity and Feasibility of Short Tests Currently Used to Measure Early Childhood Development in Large Scale Studies: Methodology and Results

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

August 2016

Cataloging-in-Publication data provided by the
Inter-American Development Bank
Felipe Herrera Library
Concurrent validity and feasibility of short tests currently used to measure early childhood development in large scale studies: methodology and results / Marta Rubio-Codina, María Caridad Araujo, Orazio, Attanasio, Sally Grantham-McGregor.
p. cm. — (IDB Working Paper Series ; 723)
Includes bibliographic references.
1. Child development-Colombia-Evaluation. 2. Early childhood education-Colombia-Evaluation. 3. Educational tests and measurements-Colombia. I. Rubio-Codina, Marta. II. Araujo, María Caridad. III. Attanasio, Orazio P. IV. Grantham-McGregor, Sally M. V. Inter-American Development Bank. Social Protection and Health Division. VI. Series.
IDB-WP-723

http://www.iadb.org

scl-sph@iadb.org

www.iadb.org/SocialProtection

Concurrent Validity and Feasibility of Short Tests Currently Used to Measure Early Childhood Development in Large Scale Studies: Methodology and Results

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

Abstract[†]

In low- and middle-income countries (LIMCs) measuring early childhood development (ECD) with standard tests in large scale surveys (i.e. evaluations of interventions) is difficult and expensive. Multi-dimensional screeners and single-domain tests ('short tests') are frequently used as alternatives. However, their validity in these circumstances is unknown. We examine the feasibility, reliability, and concurrent validity of three multi-dimensional screeners—the Ages and Stages Questionnaires (ASQ-3), the Denver Developmental Screening Test (Denver-II), the Battelle Developmental Inventory screener (BDI-2)—and two single-domain tests—the MacArthur-Bates Short-Forms (SFI and SFII) and the WHO Motor Milestones (WHO-Motor)—in 1,311 children 6-42 months in Bogota, Colombia. We compare scores on these short tests to those on the Bayley Scales of Infant and Toddler Development (Bayley-III), which we take as the 'gold standard'. The Bayley-III was given at a center by psychologists; whereas the short tests were administered in the home by interviewers, as in a survey setting. Concurrent validity of the multi-dimensional tests' cognitive, language, and fine motor scales with the corresponding Bayley-III scale is low below 19 months but increases with age, becoming moderate-to-high over 30 months. In contrast, gross motor scales' concurrence is high under 19 months and then decreases. Of the single-domain tests, the WHO-Motor has high validity with gross motor under 16 months, and the SFI and SFII expressive scales show moderate correlations with language under 30 months. Overall, the Denver-II seems the most feasible and valid multi-dimensional test and the ASQ-3 performs poorly under 31 months. By domain, gross motor development has the highest concurrence below 19 months, and language above. Results do not vary by household socio-economic status. Predictive validity investigation is nonetheless needed to further guide the choice of instruments for large scale studies.

Key Words: developmental assessment, diagnostic test, screener, concurrent validity, cognition, language, motor development, infants and toddlers, large scale studies, low- and middle-income countries.

JEL codes: J1, I1, I2, I3

[1] Social Protection and Health Division, Inter-American Development Bank, Washington DC, USA.
[2] Centre for the Evaluation of Development Policies, Institute for Fiscal Studies, London, UK.
[3] Department of Economics, University College London, London, UK.
[4] Institute of Child Health, University College London, London, UK.

[†] This working paper provides more details on the design and methodology of the study: Rubio-Codina M, Araujo MC, Attanasio O, Muñoz P, Grantham-McGregor S (2016) Concurrent Validity and Feasibility of Short Tests Currently Used to Measure Early Childhood Development in Large Scale Studies. PLoS ONE 11(8): e0160962. Any opinion, findings, and recommendations expressed herein are those of the authors and do not necessarily reflect the views of the IDB, its Board of Directors, or the countries they represent. Data collection was funded by Fund RG-T1907 from the Inter-American Development Bank (IDB). Rubio-Codina acknowledges partial financial support for research time from the Leverhulme Trust Early Career Fellowship ECF/2008/0170. Attanasio's research time was partially financed by the European Research Council (ERC) Advanced Grants 249612 and the Economic and Social Research Council (ESRC) Professorial Fellowship ES/K010700/1. We thank all the families who participated in the study; and BibloRed, Jardines Sociales del Distrito de Bogota, and aeioTU for lending us their facilities for testing. We are also grateful to all testers and interviewers, trainers (Mara Minski, Pablo Muñoz and Natalia Varela), and field coordinators (Belén Gómez, Juan Fernando Trujillo and Hanner Sánchez). We thank Stefano Banfi, Ludvig Sinander, and Camila Soares for research assistance. All errors are our own.

1. Introduction

Recent research demonstrates the importance of the early years to brain development, cognitive, language and socio-emotional development and, more generally, to human capital formation (Luby 2015; Heckman 2007). Longitudinal studies show that adversity in early childhood has sustained effects on children's development (Walker et al. 2011) and it is estimated that well over 200 million children under five years in low- and middle-income countries (LIMCs) are failing to reach their developmental potential (Grantham-McGregor et al. 2007). Interventions in early childhood can have comprehensive benefits to life outcomes (Walker et al. 2011; Gertler et al. 2014; Campbell et al. 2014) and there is an increasing global commitment to implement such interventions at large scale in LIMCs in order to promote the development of disadvantaged children. The Sustainable Development Goals (SDGs), for example, include the aim that "all girls and boys have access to quality early childhood development, care and pre-primary education so that they are ready for primary education" by 2030 (SDG 4.2) (UN General Assembly 2015).

The launching of early childhood development (ECD) interventions is nonetheless hindered by the lack of reliable and valid measures of child development that can be collected cost-effectively in large samples (Engle et al. 2007; Frongillo et al. 2014). Such measures are essential both to assess developmental levels of populations and to monitor and evaluate the effectiveness of interventions, which can inform the design of improved variants. They are also critical to estimate models of human capital accumulation that can contribute to the understanding of the process of skills formation over the life cycle, including the role of parental investments in the early years (Heckman 2007; Attanasio 2015). The need for measures of ECD outcomes is particularly pressing for children under 3 years-of-age. Hence, there is an urgency to identify readily available valid and feasible methods to assess children's development in large samples via household surveys (i.e. 'at-scale').

Multi-dimensional diagnostic tests such as the Bayley Scales of Infant Development (Bayley 1969; Bayley 2006) are considered to be the 'gold standard' to measure the developmental levels of infants and toddlers (Frongillo et al. 2014; Fernald et al. 2009; Fernandes et al. 2014). Importantly, this test has shown sensitivity to differences in ECD outcomes due to interventions in diverse contexts (Hamadani et al. 2006; Nahar et al. 2009; Attanasio et al. 2014). However, test administration is time consuming and requires highly trained professionals working in controlled environments; test kits and test administration fees are expensive; and identifying professional testers who can administer it in local languages is challenging. In addition, translation and adaptation to different languages and cultural contexts requires substantial technical skills, time, and financial resources. These reasons make the Bayley and similar diagnostic tests often infeasible for use at-scale.

As an alternative, tests designed to screen for children at risk of delay or to assess specific developmental domains (e.g. language) are increasingly used in large scale surveys and impact evaluations (Fernald et al. 2012; Macours, Schady, and Vakis 2012; Fernald and Hidrobo 2011). Although not designed for this purpose and often not validated nor standardized locally, these tests are becoming popular alternatives since they are shorter, cheaper, and easier to administer, many times being administered by regular interviewers in the children's homes, and often relying on a number of items collected by maternal report.

Nonetheless, little is known about their validity when administered at-scale not for screening but to measure levels of child development across the range of development for either research purposes or to provide population-based assessments. Two recent exceptions are the studies by Hamadani and colleagues in rural Bangladesh (Hamadani et al. 2013; Hamadani et al. 2010). The authors found moderate correlations between monthly maternal reports of age of attainment of motor milestones—primarily, walking and standing alone—and the Bayley-II Psychomotor Development Index (PDI) and low but significant associations with the Mental Development Index (MDI) at 18 months of age and with IQ at 5 years (Hamadani et al. 2013). Similarly, a language test for children 12-18 months developed locally from the MacArthur-Bates Communicative Development Inventories (Fenson et al. 2002) and administered by maternal report offered moderate concurrent validity with the Bayley-II MDI and acceptable predictive validity with IQ at age 5 years (Hamadani et al. 2010). Interestingly, maternal reports of age of walking alone and language were, respectively, as predictive of motor development or IQ at 64 months as the PDI and the MDI of the Bayley-II.

More recently, new multi-dimensional diagnostic tests for use in LMICs have become available for children 24 months or above. Examples are the INTERGROWTH-21st Project Neurodevelopmental Package (Fernandes et al. 2014) for the assessment of developmental outcomes at 24 months, or the Engle Scale, developed by the Inter-American Development Bank as part of the Regional Project on Child Development Indicators (PRIDI) (Verdisco et al. 2009) for children 24-59 months. Save the Children has developed the International Development and Early Learning Assessment (IDELA) to measure development and early learning, including early literacy and numeracy, for children 3.5-6 years (Wolf et al. 2015). Similarly, the Brookings Institution, under the Learning Metrics Taskforce Initiative, has led a number of stakeholders in the development of an instrument to measure quality of the learning environment, pre-academic and socio-emotional skills amongst 3-5 year olds, the Measuring Early Learning Quality and Outcomes (MELQO).[1] Nonetheless, these initiatives do not cover children under 2 years of age and many of the tests developed continue to be too long for use at-scale.

The current study aims to contribute to the on-going agenda on the measurement of ECD outcomes, which is rapidly attracting interest amongst researchers and practitioners alike in a variety of institutions. It is designed to investigate the extent to which a selection of multi-dimensional screeners and single-domain tests ('short tests' henceforth), are valid and feasible alternatives to diagnostic tests for the assessment of very young children at-scale. Specifically, we aim to determine the internal consistency and test-retest reliability, and the concurrent validity of five short tests administered under survey conditions to measure the developmental levels of a population-based sample of children 6-42 months in Bogota, Colombia. We also discuss their relative administration times and costs.

The short tests were selected on the basis of their current use in large scale studies in the field, and its total number was limited to avoid tiring the child. Specifically, we consider the following short tests: three multi-dimensional screeners—the Ages and Stages Questionnaires (third edition, ASQ-3) (Squires et al. 2009), the Denver Developmental Screening Test (second edition, Denver-II) (Frankenburg et al. 1990; Frankenburg et al. 1992), and the Battelle Developmental Inventory screener (second edition, BDI-2) (Newborg

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

2005); and two single-domain tests—the vocabulary checklists in the Short-Forms of the MacArthur-Bates Communicative Development Inventories I and II (SFI and SFII) (Jackson-Maldonado et al. 2003; Jackson-Maldonado, Marchman, and Fernald 2012) and the World Health Organization Gross Motor Milestones (WHO-Motor) (WHO Multicentre Growth Reference Study Group 2006; Wijnhoven et al. 2004). The latter two tests share many similarities with those used in the above Bangladeshi studies (Hamadani et al. 2010; Hamadani et al. 2013), and were included in addition to the multi-dimensional screeners since they are quicker to train and administer, and also cheaper.

To compute concurrent validity, children's developmental scores on these short tests are compared to their scores on the Bayley Scales of Infant and Toddler Development (third edition, Bayley-III) (Bayley 2006). As the gold standard, the Bayley-III was administered in ideal conditions—namely, at a center by trained psychologists. Nonetheless, and importantly to address the research questions of interest, all short tests were administered under survey conditions: in the children's home by non-specialized albeit rigorously trained interviewers.

We investigate concurrent validity of the short tests with the Bayley-III by child's age and developmental domain with a focus on cognitive, receptive and expressive language, and fine and gross motor development. Although we recognize socio-emotional development as an important developmental domain and we collected the scale, it is not included in the current analysis. This is partly because the personal-social and adaptive scales of the short tests measure slightly different constructs from the socio-emotional scale of the Bayley-III, which limits the comparison; and partly because the scale is collected by caregiver report and is reasonably quick and easy to give. We return to this issue in the next section. We also examine concurrent validity by household socio-economic status (SES) in order to explore whether some short tests are better suited for administration amongst more disadvantaged families, often less educated, and more likely targeted by government programs.

It is important to note that this study is not designed to establish the sensitivity or specificity of the screener tests in identifying high risk children. Furthermore, the number of children at risk of developmental delay in the sample is too small to allow carrying out such analyses. Rather, we are interested in examining the ability of the short tests to measure child development across the range of developmental levels in our study population, representative of low- and lower-middle income groups in a typical large city in Latin America. The aim is to identify feasible and easy-to-use readily available instruments for the assessment of very young children in large scale studies and in contexts different for those for which the tools are developed, thus guiding the choice of instruments for research purposes (for example, in program evaluations) and/or population-based assessments.

The paper is structured as follows. Section 2 describes the study design and data collection strategy. It also includes a description on the child assessments administered and the final sample of analysis. Section 3 presents the strategy used in the empirical analysis. Section 4 presents results and Section 5 discusses issues related to test choice for use at-scale in the light of the analysis and concludes.

2. Study Design and Data

2.1. Participants and Data Collection Strategy

Bogota is divided into six socio-economic strata, denominated 'sectors' ('*estratos*'), based on location and quality of housing and infrastructure. The study enrolled a representative sample of children aged 6-42 months, randomly selected from the poorest three sectors and stratified by age and sector. These three sectors account for 85% of the city's population and comprise low- and lower-middle-income households.[2] While we had originally included Sector 4 (middle-income) in the study design, it was subsequently dropped due to the difficulties in contacting and obtaining participation consent amongst households from this sector, who often live in restricted access apartment blocks and compounds. Mistrust was one of the main reasons behind the high participation refusals. Regarding the ages of the children included in the study, we set the lower-age limit to 6 months since earlier measurements have lower predictive ability of future development and given cost considerations.[3] The upper-age limit was determined by the use of the Bayley-III, which is designed to assess children up to 42 months.

Data were collected between March and August 2011. Prior to data collection, we did not have access to administrative records with both date of birth and home address, including socio-economic sector. Yet, our sample included very young children, and had to be representative by socio-economic sector and balanced by age. In addition, and in order to minimize seasonality or tester learning or fatigue effects, it was important to ensure that children of all ages and all socio-economic sectors would be tested in similar proportions over the data collection period. Complying with all of these requisites posed a substantial logistical challenge and required following a well-defined sampling and data collection strategy, which we strictly implemented in three stages.

Firstly, neighborhoods (and blocks within them) were selected using the proportion of women in fertile age as weights (probability design). Once selected, we visited by door-to-door census all households in a block to identify those with children aged 6-42 months. These activities were carried out by a team of interviewers, who were exclusively devoted to identifying the study sample. Children with learning disabilities (one child) and twins (one pair) were excluded from the study for practical reasons. Similarly, in households with more than one child in the relevant age group for the study (four cases), one was randomly selected and included. The remaining eligible children were stratified by age category and 80% (per block and age group) were randomly drawn for study inclusion.

Next, all included children in a block were randomly assigned to one of eight non-specialized trained interviewers, who visited their homes to collect the short tests and a household survey. The latter included basic household socio-economic information (such as demographic composition, education, and employment for all household members, dwelling characteristics, and assets); the child's health history (birth weight, gestational age, etc.); data on formal and informal childcare arrangements; as well as the quality of the home environment using UNICEF's Family Care Indicators (FCI) (Frongillo, Sywulka, and Kariger 2003). Specifically, we recorded, by observation, the number of books for adults,

[2] Neighborhoods in the bottom two sectors are typically considered poor, while those in the third sector are considered lower-middle class. It should be stressed, however, that there is substantial heterogeneity in household socio-economic background within sector, especially in neighborhoods that have recently developed (see Rubio-Codina et al. 2015).

[3] It is also likely that including younger children would have limited household willingness or ability to participate in the study, particularly given that the Bayley-III and anthropometric measurements were collected outside of the child homes.

newspapers/magazines, and the toys the child usually played with by type; and by caregiver report, the play activities the child and an adult engaged in over the week prior to the survey.

In a third and final stage, the Bayley-III test was administered by trained psychologists (testers) in the public library or public childcare center closest to the child's home.[4] This ensured all Bayley-III assessments took place in a similar environment that satisfied standard testing requirements (quietness, light, space, ventilation), thus facilitating the child's concentration on the test and optimizing testing time. On average, children were tested on the Bayley-III, 5-6 days after the short tests assessments (78% within a week and 94% within two weeks) and the testers were blind to children's performance on the short tests. After completion of the Bayley-III assessment, the testers collected height and weight of both mother and child following WHO guidelines (WHO 1983). As a token of gratitude for their participation in the study, tested children were offered a set of picture books and nutritional supplements (vitamins and minerals) for daily consumption over 3 months. Similarly, we gave the mother feedback on her child's performance in the test, a set of brochures on parenting, and $10,000 pesos (about $5.6 US) to cover transportation costs to the testing site.

To increase the number of tests examined, and minimize test weariness, children were randomly assigned to one of two batteries of short tests. Battery A included the ASQ-3, the Denver-II, and for children between 8 and 30 months the SFI or the SFII, depending on the child's age. Battery B comprised the BDI-2 and for children 6-15 months the WHO-Motor. The short tests were administered in the order listed within the battery and after the first section of the household survey, once rapport with the caregiver had been established. The administration of both batteries took similar amounts of time, the length of the total household visit (household survey + short tests) being no more than 2-2.5 hours. This allowed completion of 2-3 household visits per interviewer per day, the average number of daily interviews increasing as data collection activities progressed. Similarly, each tester administered between 2 and 3 Bayley-III tests a day. Between 2.5 and 5% of the sessions, either in the home or the center, had to be rescheduled because the child was too sick or fussy to be tested.

All measurements (short tests and Bayley-III) took place in the presence of the main caregiver—the mother in 85-89% of the cases, the father in another 3-5% of cases. For the remaining cases, main caregivers were often other relatives. Caregivers responded to test items when appropriate. For this reason, and to ensure the child felt she was in the company of someone familiar and supportive during the assessment, we requested the person accompanying the child to be older than 15 years and to typically spend at least five hours taking care of the child over a minimum of five days a week.

Figure 1 summarizes the study design and stages, lists all tests administered by battery, and reports the number of participants at each stage and test. We strictly monitored the ages and sector of all children enrolled throughout the process in order to guarantee a final well-balanced sample. Moreover, data collection was organized such that all interviewers and testers assessed similar numbers of children in each sector and age group, in a uniform manner over the six months of field activities. This was important so as to minimize potential

[4] This was through a partnership with the local network of public libraries *BibloRed* and the public child care centers *Jardines Sociales.* In return for lending us their facilities, we offered workshops on parenting and child rearing practices to the centres'/libraries' staff and parents.

measurement biases due to: (i) child socio-economic status (e.g. the tester scores children from different backgrounds differently to compensate for perceived disadvantages); (ii) child age (e.g. the tester finds it easier to test older children); (iii) seasonality (e.g. measurements are less accurate because they are administered faster near holiday periods or long weekends); and (iv) tester learning effects or tester fatigue (e.g. testing is more accurate during the middle months of data collection when the testers have had enough practice, but are not too tired of administering the same test). In other words, we wanted to avoid that any patterns observed in the data by age or by socio-economic sector were due to any of the potential sources of bias listed.

The ethical committee at the Instituto de Ortopedia Infantil Roosevelt in Bogota reviewed the study protocols and considered them to be fully compliant with required ethical practice. Written informed consent to participate in the study was obtained from parents on behalf of the children enrolled. Further details on the sample selection and data collection procedures are provided in Rubio-Codina et al. (2015) and Rubio-Codina, Attanasio and Grantham-McGregor (2016).

2.2. Child Assessments

The first and second columns in Table 1 detail the test and scales (i.e. developmental domains) we administered in the study. It is worth noting that while three of the short tests—namely, the ASQ-3, the Denver-II and the BDI-2—cover multiple dimensions, the WHO-Motor and the SFI and SFII are single-domain tests, focusing only on gross motor and language development, respectively. Next to each scale, we report the total number of items in the test in parentheses and the average number of items assessed per child in the study sample in brackets. For those tests with start and stop rules, the latter number is a function of the child's age and ability and hence, in these cases, the two values do not coincide.

The following columns report the age range covered by the test and in the study—note that not all tests apply to children across the entire study age range; and some other test characteristics, including the cost of purchase of the test kit (excluding shipping and custom fees) and the per child administration fee, the time to administer as reported by the test publisher, and the average administration and training time in the study.[5] The final two columns present the trainers' assessment on difficulty to train and difficulty to administer each test.

Even if most tests were available in Spanish, some had to be translated either partly or fully. Moreover, piloting of the existing (official) Spanish version or the translated version suggested minor wording and phrasing modifications to better reflect Colombian Spanish; as well as the contextualization of a few images. We list all modifications, as well as the publisher website for each test in Appendix I. We describe them in more detail next.

2.2.1. Criterion Measure: Bayley Scales of Infant and Toddler Development, third edition (Bayley-III)

The Bayley-III (Bayley 2006) is a tester-administered diagnostic test consisting of the following scales:

[5] Total time to administer the test was recorded by the trainer during those assessments that were supervised (approximately, 5% of the sample).

(i) Cognitive Scale. It primarily requires non-verbal responses from the child and measures learning processes, problem solving, attention, counting and classification, and playing skills, amongst other constructs.
(ii) Language Scale. It comprises the language receptive and expressive subscales. The first measures the child's ability to respond to stimulus in the environment, words, and requests. The latter assesses the child's vocalizations and use of words and sentences.
(iii) Motor Scale. It includes the fine motor subscale, measuring hand-and-fingers and hand-and-eye coordination, and the gross motor subscale, which measures the child's body control and movement of the torso and extremities.
(iv) Socio-emotional Scale, which uses the Greenspan Social-Emotional Growth Chart (Greenspan 2004). It measures social and emotional milestones, such as self-regulation, attention, how the child relates and interacts with familiar and non-familiar people, and other temperament and social aspects.
(v) Adaptive Behavior Questionnaires, which use the Parent/Primary Caregiver Form of the Adaptive Behavior Assessment System (second edition, ABAS-II) (Harrison and Oakland 2003). These comprise ten subscales measuring daily functional abilities of children 0-5 years.[6]

The scales are administered and scored independently, producing domain-specific assessments. The cognitive, language and motor scales are assessed by direct observation of the child's abilities in items arranged in increasing order of difficulty. Basal and ceiling rules determine the starting and stopping points. The child scores 1 for each item correctly executed and 0 otherwise. The raw score is the sum of correct responses, including non-administered items preceding the basal.

As indicated earlier, the focus of the current study is on cognitive, language and motor development. The socio-emotional scale comprises 35 5-point rating questions responded by the caregiver, hence being reasonably quick and easy to give. But more importantly, since the personal-social and adaptive scales of the short tests are more related to self-care and self-direction, they measure slightly different constructs from the socio-emotional scale of the Bayley-III. As this makes comparisons between scales less straightforward, we do not include socio-emotional development in the analysis. Regarding the adaptive behavior questionnaires, only two subscales in the ABAS-II were collected on a subsample of children, given time constraints and the age- and context-inappropriateness of some of the items in many of the other subscales. Therefore, these scales, also administered by parental report, are not included in the analysis either.

The Bayley-III requires the test to be administered by child development professionals, such as psychologists and educators, after undertaking a rigorous training. Administration times range between 30 to 90 minutes, depending upon the age of child. In our case, administration of the cognitive, language, motor, and socio-emotional scales took 83 minutes on average, and varied from 40 to 150 minutes depending on the child's characteristics (age, interest, attention, etc.). In fact, time to test strongly increases with age in a linear fashion for children younger than 24 months, and plateaus thereafter. Average testing time is thus 77 minutes for children younger than 24 months and 93 minutes for older children.

[6] The ten areas covered are: communication, community use, functional pre-academics, home living, health and safety, leisure, self-care, self-direction, social, and motor.

The test kit costs $1,050 US and includes a stimulus book, a picture book, a set of manipulatives (dolls, balls, ducks, pegboards, form boards, puzzles, blocks, etc.), the technical manual and administration manuals, and 25 record forms for each scale. Additional record forms can be bought for each individual scale or for various scales. Each additional record form costs $9.34 US for all scales, or $5.02 US for the cognitive, language, and motor scales. This unitary cost per record form is equivalent to a per child administration fee, since the publisher requests that a form is bought per each individual child. In a large scale study, this can easily result into near-to-prohibitive administration costs. Moreover, there are some additional materials required for administration that are not included in the test kit, such as scissors, tape, pencils, paper, a stopwatch, and a set of steps of specific dimensions, which are required for the administration of the gross motor scale. Purchase of the test is limited to individuals with specific backgrounds (e.g. doctorate degree in psychology, education or closely related field) or a certification to practice/full membership in specified professional organizations, and training in test administration and expertise in test interpretation.

The test became available in Spanish in mid-2015. This meant we had to translate the English version of the test manual and record forms to Colombian Spanish, and back-translate them to English.

2.2.2. Short Test for Validation in Battery A

2.2.2.1. Ages and Stages Questionnaire, third edition (ASQ-3)

The ASQ-3 (Squires et al. 2009) is a screening tool for children 1-66 months, comprising 21 age-specific caregiver-completed questionnaires. Each questionnaire assesses child competence in five domains (scales)—problem solving (or cognition), communication, fine motor, gross motor, and personal-social—with six items in each.

As a screener test, it is designed to identify children at risk of developmental delay and therefore is more sensitive for the measurement of development at the lower end of the distribution of skills. However, our aim was to assess the test's suitability for use in an intervention evaluation setting and hence its ability to measure child development across the entire distribution of skills, including children developing above average. Hence, we modified the administration and gave the first three new items from subsequent questionnaires whenever the child attained the maximum score in a scale. This increased the variability of developmental abilities captured by the test and reduced the number of children on the test ceiling by 10.5-15.5% to levels of 1.7-4.8%, depending on the domain. Moreover, because of the low education levels of some caregivers, items were given by interview, as opposed to having caregivers complete the questionnaires on their own. In addition, the interviewer would administer an item directly to the child if the caregiver could not provide an answer to the item or if the item wording implied one should test the child to see if she could perform it or not. The ASQ-3 manual encourages trying out items with the child, especially if the test is completed with support from trained (para-)professionals. Similar adaptations of the ASQ to the ones described have been used previously in other studies in middle- or low-income countries (Fernald et al. 2012).

A score is calculated for each scale and questionnaire. Answer options 'yes', 'sometimes' or 'not yet' are allocated 10, 5 or 0 points, respectively, and totaled. Missing items are replaced with the scale average (1.2% of children in the sample). However, if more than two items are missing in a scale, the scale is not scored (0.3% of cases).

Given these modifications to standard administration protocols, testing time increased to almost 20 minutes on average, from the 10-15 minutes reported in the publisher's website. The test is available in Spanish and the Starter Kit, including photocopiable print masters of the questionnaires and scoring sheets in Spanish, a CD-ROM with printable PDF questionnaires, and a user's guide in English, costs $275 US. A Materials Kit, including approximately the 20 toys, books and other manipulatives, designed to encourage the child's participation in the test, and to support effective and accurate administration of the questionnaires is available for $295 US. While the use of materials is needed if items are directly tested on the child, it is not compulsory to use the materials from the kit and these can be replaced with manipulatives of similar characteristics. Note however that, when assessing development at large scale, particularly for the evaluation of interventions, it is critical to standardize administration protocols to ensure that differences in developmental levels are not due to tester idiosyncrasy in the administration or in the scoring of the test. To this end, having a common standardized set of materials for all testers is recommended.

2.2.2.2. Denver Developmental Screening Test, second edition (Denver-II)

The Denver-II (Frankenburg et al. 1990; Frankenburg et al. 1992) is a screener test designed for use by a clinician or early childhood professional to monitor the development of children 0-6 years and identify significant deviations in development. Most items (68%) require the examiner's actual observation of the child's behavior or performance in the item, but some can be assessed by parental report—particularly in the personal-social (76%) and language scales (38%).

It comprises four scales—language, fine-motor/adaptive, gross motor, and personal-social—which are administered and scored independently. Basal and ceiling rules around an age line determine the test items to be given to each child, which are arranged in increasing order of difficulty. The child scores 'pass' for each item correctly executed or positive parental report; and 'fail' otherwise. 'No opportunity' (the respondent does not have a chance to observe the behavior) and 'refusal' (child refuses to attempt the item) are valid options for caregiver-reported and administered items, respectively. Children with at least one 'refusal' item to the left of the age line are considered untestable and the scale is not scored (0.5% in the sample). The test categorizes children as 'normal' or 'suspect' depending on their performance relative to children in the reference population. However, to compute concurrent validity we required a continuous score that we could correlate to the Bayley-III score for each scale. Therefore, for each domain, we constructed a 'raw' score by scoring 'pass' as 1, 'fail' as 0, and adding up the sum of responses, crediting non-administered items preceding the basal. We scored 'no opportunity' and 'refusals' with a 0.

Administration times are reported to be around 15 to 20 minutes, whereas we took 27 minutes on average. It is likely that the administration times reported by the test developers assume administration by a pediatrician or similar profile. The test kit costed $200 US at the time of purchase for the current study, and included the test technical manual (in English), trainers manual in English, the record forms in Spanish, a DVD with administration instructions, and a small bag with the manipulatives required for administration of the test (with the exception of blank sheets of paper). Each additional record form costed $0.45 US. Since 2015, however, the publisher has discontinued the test, even if the manuals and report forms are still available for download from their website. The website also includes a picture

showing all the manipulatives (toys and other materials) needed for administration, but these are no longer available for purchase.

2.2.2.3. MacArthur-Bates Communicative Development Inventories I and II, Short-Form versions (SFI and SFII)

The Spanish-language MacArthur-Bates Communicative Development Inventories I and II (S-CDIs) (Jackson-Maldonado et al. 2003) are well-established parent report tools for assessing the language development of Spanish-speaking children 8-18 and 16-30 months, respectively. Short-form versions of the S-CDIs, the SFI and SFII, were developed as alternatives for screening purposes or applications requiring a less-demanding instrument, and were validated in Mexico (Jackson-Maldonado, Marchman, and Fernald 2012). We used the vocabulary checklist in the SFI to asses receptive and expressive language—number of words the child 'understands' and number of words the child 'understands and says', respectively—for children 8-18 months; and in the SFII to measure expressive language—this is to say, the number of words the child 'says'—for children 19-30 months.[7]

Raw scores are computed by counting the number of words the child 'understands', 'understands and says', or 'says', depending on the checklist. In the SFI the score for word comprehension must always be equal to or greater than that for word production. Items left blank are not counted.

All that was required for administration were the vocabulary lists. These are available from the CDI Advisory Board at Stanford and the cost per use is determined on a case-by-case basis, depending on the use of the test.[8] The complete S-CDIs, including the manuals, are available for purchase at a cost of $90 US, and $1 US for each additional record form. Since these forms were designed and validated for Mexico, some words may need to be replaced for use in different Spanish speaking countries to ensure linguistic and functional equivalence of the word item—for example, in Colombia, '*punta*' is the common word for 'clavo' (nail). We took about 8 minutes on average to complete each vocabulary checklist by interview to the caregiver.

2.2.3. Short Test for Validation in Battery B

2.2.3.1. Battelle Developmental Inventory screener, second edition (BDI-2)

The BDI-2 screener (Newborg 2005) was developed to identify risk of developmental delay for children under eight years of age. It comprises five scales—cognitive, communication, motor (combining fine and gross motor items), personal-social, and adaptive skills—which are administered and scored independently. The test indicates the preferred procedure to use in the administration of each item: (i) *structured administration*, directly testing the item on the child; (ii) *observation* of the child's abilities for an extended period of time (usually during the testing session); and (iii) *interview* with the caregiver.[9]

Items are arranged in increasing order of difficulty and basal and ceiling rules determine the number of items each child is tested on. The child scores 0 for each item that she cannot complete, 1 for each item that she can complete partly, and 2 for each item that she can

[7] A short-form for children 30-37 months was under development at the time of this study. However, we did not learn about it until after we had completed data collection.
[8] http://mb-cdi.stanford.edu/board.html.
[9] In the 31 items in which long observation periods (days or weeks) were required, we substituted "observation" by "interview" as the preferred administration procedure.

complete fully. The raw score is the sum of all responses, crediting with a 2 all items preceding the basal. Missing items are scored with a 0 (1.8% of the cases).

The manual encourages that testers have college-level training, preferably in psychology or related disciplines, although the possibility of administration by non-professionals after rigorous and supervised training on the test and on measuring children is also accepted. Test administration time is reported to be between 10 to 30 minutes in the manual. On average, we took 59 minutes to administer the full test, which is substantially longer. The administration times in the manual are likely to assume administration by professionals with backgrounds in relevant disciplines and more familiarity in the assessment of children. In any event, 10 minutes (i.e. the lower bound indicated in the test manual) seems too little time to administer 9 items on 5 scales, on average. In the study, testing time increased with age for younger children and up to 24 months.

The BID-2 screener kit in Spanish costs $405.70 US and includes the examiner's manual, the test item book, a set of presentation cards, the stimulus book, a pack of 30 record forms, and the manipulatives needed to administer the screening test. The cost of each additional record form is $3.08 US. Materials required some adaptation and translation as some of the content in the Spanish kit was in English. Specifically, the test item book (manual), which includes the specific instructions for accurate item administration and scoring, had to be translated from English to Spanish. Similarly, the text in the picture (story) book had to be translated. To purchase and use the tests, the publisher requires the provision of relevant qualifications, which are consistent with sound professional practice.

2.2.3.2. World Health Organization Gross Motor Milestones (WHO-Motor)

The WHO-Motor (WHO Multicentre Growth Reference Study Group 2006; Wijnhoven et al. 2004) includes six milestones to assess the gross motor development of children 6-18 months. Analysis was however limited to children 6-15.9 months, since 91.9% of those older, attained all milestones. All milestones were given by direct administration and we did not collect parental records on date (or age) of achievement of each milestone.

Since the test does not provide indications on how to compute a raw score, we added up the number of milestones the child was observed to perform, crediting earlier milestones. We dropped 3 children (1.4%) with inconsistent or missing data.

The test is available for free in English from the WHO. We translated the report form and the administration instructions to Spanish.

2.2.4. A Short Note on Prematurity

We did not adjust for prematurity prior to the administration of any of the tests. Instead, we started premature children at the corresponding unadjusted start point and had them go back to earlier (easier) items as required given their developmental level. While this may increase testing time, it deals with potential inaccurate caregiver reports on gestational age.[10] The only exception to this rule was the ASQ-3, where we followed the test manual protocols given the questionnaires are age-specific.

[10] In fact, we observe 9% mismatches (over 50% of those reported as premature) in reported weeks of gestation between the household and Bayley-III surveys.

2.2.5. Tester and Interviewer Profile and Training

Six female psychology graduates, some with previous experience assessing children, were trained on the Bayley-III for six weeks, including 20-25 practice administrations per tester on children across the age range. None of them knew the test. They were also trained on how to collect height and weight for 2.5 days (10-12 practices). Eight female interviewers, with no university education and no prior experience testing children, were trained on the short tests in either battery A or battery B for 6-7 weeks, including the training on the household survey. On average, they carried out about 20 practice administrations on each short test (they often administered the entire battery of short tests, A or B).

Practice testing for testers and interviewers occurred in pairs and inter-observer reliabilities (degree of agreement) between trainee-trainer and between each pair of testers/interviewers were collected. To ensure a standardized administration, it is advisable to continue practice testing until inter-observer reliabilities are satisfactory (intra-class correlations, *ICCs*>0.9) on each scale and test. Table 1 reports our best approximation to the number of days required to train on each test. This is based on our experience in this and other studies, but is subject to vary as a function of the qualifications and previous experience of the trainees. The number of practices also depends on the complexity of the test and tends to increase with the number of items tested on the child or scored by observation, as opposed to by caregiver report. Consistently, tests and scales with more items by caregiver report are easier to train and administer, as assessed by the trainers. We worked with three trainers—one for each set of tests: Bayley-III, short tests in battery A, and short tests in battery B. All trainers had master degrees in Psychology.

During data collection, 5% of the field assessments were observed and scored by the relevant trainer and inter-observer reliabilities were computed. The trainer gave corrective feedback whenever appropriate. The agreement between interviewer/tester and trainer scores during these tests was high (*ICC*s mean =0.95), hence indicating that testing quality was sustained throughout data collection.

2.3. Analysis Sample

Figure 1 graphically depicts the flow of study participants, and final number of subjects in each assessment. Data were collected on a sample of 1,533 children aged 6-42 months in 497 blocks, mostly in sectors 1-3 of Bogota.[11] The Bayley-III test was however administered to 1,330 (86.8%) of the children for whom we have a household survey and the relevant short test(s). The remaining 13.2% of children that were not assessed on the Bayley-III were more likely to attend a child care center, have younger mothers, and live in households with older children and/or no elders. This suggests that mothers without alternative forms of care may have faced difficulties finding the time to take the child to the test. Rubio-Codina et al. (2015) show that the resulting sample of children with Bayley-III data remained representative of household SES.

Of the 1,330 children with Bayley-III, 4 (0.3%) did not complete the test and 15 (1.1%) scored <70 in any of the Bayley-III composite scales and were dropped. The remaining 1,311 children with complete and consistent Bayley-III data constitute our sample of

[11] We assessed 403 children in 134 blocks in Sector 1, 459 children in 159 blocks in Sector 2, 457 children in 199 blocks in Sector 3, and 12 children in five blocks in Sector 4.

analysis. They were administered the age-corresponding short tests in battery A (n_A =676) or battery B (n_B =635), as decided by random assignment stratifying by age and socio-economic sector.

Table 2 presents summary statistics for a selection of characteristics of these children, their parents, and their households, by battery of short tests. For each variable, the last column reports the p-value of the difference of the mean between the two batteries. Around 15% of the children in the sample are premature (gestational age <37 weeks) and 17-18% are stunted (z-score height-for-age <-2 standard deviations (SDs) of the median WHO growth reference (WHO 2006)).[12] Mothers are 26-27 years old and have slightly over 10 years of education on average, with 30-31% of the mothers with more than secondary education, and 50-54% of them reported having a job, either paid or unpaid. Fathers have slightly fewer years of education than mothers, and do not live with the child in 32-34% of cases.

Both batteries are well-balanced by child's age and sex (Panel I), and by socio-economic sector (strata, Panel III), as expected by design—this is to say, by the random assignment of children within age group and socio-economic sector to each battery. Whilst other child and parental characteristics are also well-balanced between batteries, households in battery B seem to be slightly richer, as indicated by the significantly higher wealth index (p =0.013, Panel III). This is consistent with the significantly higher 0.6 years of education reported for fathers in battery B (p =0.026, Panel II) and is possibly related to the large heterogeneity in household wealth within sector, documented in Rubio-Codina et al. (2015). The household wealth index is constructed using polychoric principal component analysis on a set of household assets and dwelling characteristics, following Rubio-Codina et al. (2015).[13] The quality of the home environment, as measured by the number of varieties of play activities and of play materials in the home collected using the FCI, is also similar amongst households in either battery.[14] Importantly, children in batteries A and B have comparable developmental levels, as assessed by the Bayley-III (Panel IV).

3. Statistical Analysis

3.1. Internal Standardization of Scores

For each scale, we construct continuous raw scores following the instructions in the tests manuals, and as indicated above. Since the Denver-II has no raw score, we add items passed to items preceding the basal level, following general scoring principles. In the same spirit, we construct the WHO-Motor raw score by adding all milestones the child passed, and crediting earlier milestones. Hence, raw scores increase with age for all tests by construction, except for the ASQ-3 which had age-specific questionnaires. Table A1 in Appendix III reports raw scores for the short tests for all children 6-42 months and by 12-

[12] All measurements were converted to WHO growth reference standard scores using the WHO Anthro software (version 3.2.2, 2011).

[13] Variables included were: car, fridge, microwave, washing machine, boiler, computer, smartphone, flat TV, home theatre, DVD, stereo, games console, internet, garage, whether the household shares the kitchen with other households, whether the household shares the bathroom, has more than one bathroom, has quality floors (tiles, carpet or wood as opposed to gravel, cement or dirt), has external windows, and the crowding index. The first principal component explained 43.09% of the total variance, the second component explaining an additional 8.03%.

[14] The number of varieties of play materials is the sum of indicators for: toys that make/play music; toys/objects meant for stacking/constructing/building; things for drawing/writing/coloring/painting; toys for moving around; toys to play pretend games; picture books and drawing books for children; and toys for learning shapes and colors). The number of varieties of play activities is the sum of indicators for: reading books/looking at picture books; telling stories to child; singing songs with child; playing with child with toys; spending time with child scribbling/drawing/coloring; spending time with child naming things/counting; and taking child outside for a walk.

months-of-age groups—this is, for children 6-18 months, 19-30 months, and 31-42 months. The age gradient in mean raw scores is apparent from the values in the table. For the ASQ-3, scores fluctuate in an arbitrary manner showing a certain tendency to increase with age, particularly for children in the oldest age group.

To correct for the age effect, raw scores need to be standardized over age. However, neither the Bayley-III nor any of the short tests have been standardized (normed) for Colombia before. Moreover, the Bayley-III composite scores were shown to vary by age in unusual ways in this sample, with decreasing composite scores for cognition with age, increasing for motor development and U-shaped for language (Rubio-Codina et al. 2015). In addition, the SDs are smaller than the expected 15 points in the standardized population and decreased with age, particularly for cognition. These patterns, also shown in Table A1 in Appendix III, are suggestive of the unsuitability of the Bayley-III external norms (derived from a sample representative of the US population) for our sample. Similar conclusions can be drawn from the fluctuations observed in the ASQ-3 means and SDs just described. Therefore, as commonly done with developing country data, we internally standardize scores over age. Unlike using norms from the reference populations for each test (external standardization), internally standardizing scores in the same manner for all tests has the advantage that it handles age effects consistently across tests, thus facilitating test comparisons.

Often, the internal standardization is done by dividing the sample into the smallest possible age groups—ideally monthly, given how sensitive developmental milestones are to age in the early years—but guaranteeing enough observations per group, and computing *z*-scores within age groups (see Fernald et al. 2011, for example). We follow this approach but compute internal *z*-scores in a more flexible manner while taking into account our limited sample size. In particular, instead of using months-of-age-specific means and SDs, we estimate age-conditional means and SDs using non-parametric methods as described in Appendix II. This procedure is less sensitive to outliers and small sample sizes within age category, and more closely replicates the way in which the tests would compute external scores (since it is completely non-parametric).[15] Note that, in order to correct for tester/interviewer idiosyncrasies in the administration and scoring of the tests, we internally standardize over age the *residuals* of the raw scores, net of testers'/interviewers' effects, as opposed to the raw scores directly.

3.2. Investigation of Test Reliability, Validity, and Feasibility

After providing empirical evidence to support the validity of the Bayley-III as our gold standard in this exercise, we move to the discussion of the reliability, validity, and feasibility of the short tests.

We first examine the short tests' test-retest and internal consistency reliabilities by scale. Reliability is the degree to which an assessment tool produces stable and consistent results. Test-retest reliability is a measure of test stability over time and is obtained by computing the *ICC* of the scores in a scale from two different assessments administered on the same child by the same interviewer/tester but within a few days apart, often between one and two weeks. Internal consistency reliability explores the extent to which items in a scale (or test)

[15] For example, in the case of the Bayley-III composite scores are a nonlinear function of raw scores. Specifically, (i) items administered in younger ages are given more weight, and (ii) months-of-age are lumped together until the age of 36 months, and in larger intervals thereafter.

measure the same underlying construct (domain or skill). We compute it using Cronbach's alpha (□) on all items in a scale for all children in the sample (6-42 months) and by 12-months-of-age groups.[16] Higher reliabilities are an indication of higher performance of a test in a given population. The analysis of reliability is of particular importance whenever a test is administered in a population different from that for which it was designed, especially if the language or content in the items has been modified to ensure comprehension and item functional equivalence. We also explore the extent to which the scales in a test are correlated amongst each other, which is an indication of congruence between scales and speaks to the interrelatedness of developmental domains.

Next, we investigate validity. Validity refers to how well a test measures what it is supposed to measure. It is generally considered the most important element in psychological testing because it concerns the meaning placed on test results. In this study, we will focus on criterion validity—this is the correlation of test results with another criterion of interest. Criterion validity can be concurrent or predictive, depending on whether it concerns the prediction of current or future test performance. Given our data is cross-sectional, we can only study concurrent criterion validity, to which we will refer as concurrent validity hereon.

We start by investigating concurrent validity between scores in each test scale and a set of variables theoretically related to child development by computing Pearson correlations (*r*) by domain (scale). Variables considered include maternal education, the household wealth index, the FCI scores for play activities and play materials in the home, and two indicators for prematurity and stunting, respectively. We then turn to the core of our analysis, which is the examination of concurrent validity between the short tests and the Bayley-III (our established gold standard criterion). We do this by computing Pearson correlations (*r*) by domain and by age groups. Since all correlations use the internally standardized scores—i.e. internally standardized over age after removing tester/interviewer effects—this is equivalent to computing partial correlations controlling for testers/interviewers and age flexibly. *P* values for the correlations were computed using bootstrapping methods, with 1000 replications and clustering by age and sector (Efron 1982). Following Evans (1996), we classify Pearson correlations as low (*r*=0.20-0.39), moderate (*r*=0.40-0.59), and high (*r*=0.60-0.79) throughout the presentation of results and discussion.

We also use bootstrapping methods to compare the size of the correlations of each of the short tests with the Bayley-III by age group, and identify those that are statistically significant from each other. For example, we test whether the correlation between the BDI-2 and the Bayley-III cognitive scales is significantly larger or smaller than the correlation between the ASQ-3 and the Bayley-III cognitive scales. We do this for each pair of correlations (within an age group) that exhibit large enough differences to merit testing for statistical significance.

We carry out a number of robustness checks. We start by investigating robustness of the concurrent validity results to the use of the Bayley-III composite scores (external standardization) and to parametric standardization methods. Next, we repeat the analysis using the original 6-item version of the ASQ-3, controlling for prematurity before standardizing scores, and by further dividing the sample by 6-months-of-age groups. Finally, we compute Spearman rank-order correlations to examine the monotonicity of the

[16] Since the ASQ-3 has age-specific questionnaires, items vary from questionnaire to questionnaire. Hence, we compute the internal consistency for each questionnaire independently and average the □'s for all questionnaires covering the age group of interest.

relationship between any two scores, as opposed to the linear correlation. Given space limitations, some of these robustness tests are included in Appendix III and some are available upon request, as will be further detailed below.

We then investigate whether concurrence between the short tests and the Bayley-III varies with households SES. Specifically, we replicate the analysis of concurrent validity by domain independently for households in the bottom and top 25% of the household wealth distribution. As before, we test whether the size of the correlation between the 25% poorest and 25% richest households in the sample is significantly different (statistically speaking) using bootstrapping methods. We focus on concurrent validity for matching developmental domains only. Moreover, we work with the complete sample of children (ages 6-42 months) given the limited sample sizes and to avoid failing to reject the null of no differences in the correlations by household SES due to a lack of power.

Lastly, we discuss feasibility of test administration. This relates to all costs involved in the purchase and administration of the test, which include those incurred during test adaptation and training.

4. Results

4.1. The Bayley-III as the Gold Standard

Table A1 in Appendix III shows that mean composite scores of the Bayley-III are in the normal range, despite displaying some unexpected relationships with age. The SDs are lower than expected and also decreasing with age, particularly for cognition. As discussed in the previous section, this further justifies the pertinence of the internal standardization.

The first two panels in Table 3 show test-retest and internal consistency reliabilities for the Bayley-III raw and composite scores. Test-retest reliabilities on 20 children after 6-19 days (median of 8 days) are very good, all *ICCs* ≥0.96, which suggests very good test stability of the Bayley-III over time. Similarly, internal consistency seems to be very good across the age ranges for all domains (scales), even if we do not observe a clear pattern with age. Both reliability measures are higher for the Bayley-III than for any of the short tests, which further supports its use as the gold standard.

Furthermore, the correlations of the Bayley-III scales among each other for the entire sample of children 6-42 (first panel in Table 5) and by age group (first panel in Table 6) are similar to those reported in the test manual (Bayley 2006). They show a slight tendency to increase with age.

4.2. Reliability of the Short Tests

The third to ninth panels in Table 3 show test-retest and internal consistency reliability for the short tests. Internal consistency is first reported for all children in the sample and then by age group.

Test-retest reliabilities after 2-11 days (median of 8 days) are available for 12 children for the short tests in battery A; and after 5-11 days (median of 7 days) for 11 children for those in battery B. Despite the small sample sizes, the values are generally satisfactory (*ICCs* ≥0.7), indicating stable measurements over time for most of scales. The only exceptions are the

fine motor scale in the 9-item and 6-item versions of the ASQ-3 (both r =0.37) and the gross motor (r =0.53) and personal-social (r =0.49) scales in the Denver-II.

When considering all study children (second column in Table 3, 6-42 months), Cronbach's □□□are generally good and above the desired cut-off of 0.7, which suggests good internal consistency. The only exceptions are for the 9-item and 6-item versions of the ASQ-3 which had lower values in all scales except gross motor in the 9-item version. As observed with the Bayley-III, the internal consistency of the short tests does not seem to follow any consistent pattern with age, although values are generally lower within age groups than when the entire sample of children is considered. By age group, most of the ASQ-3 and BDI-2 scales, especially for children in the middle age group in the case of the BDI-2, show poor or very poor internal consistency. We have also investigated but could not identify any specific items that were particularly problematic—this is to say, items that do not measure the same underlying construct and hence could explain the poor internal consistency of some scales.

Note that the internal consistency of the 6-item version of the ASQ-3 is substantially lower than that of the 9-item version of the ASQ-3 for every scale and age group. In fact, all but one scale show low internal consistency, with values □□≤0.55. This further supports our choice to use the 9-item version and as such, all results reported henceforth will use the 9-item version.[17]

The correlations amongst the scales in the ASQ-3, Denver-II and BDI-2 tests (available upon request) are usually lower than those observed amongst scales in the Bayley-III, both on the entire sample and by age group, indicating yet again the Bayley-III preeminence as a gold standard. For the ASQ-3, values are also slightly lower to those reported in the test manual, which may be related to the adaptations and modifications made to the administration of the test. Note that the comparison is unfair since the correlations reported in the manual cover ages 1-66 months whereas the ones we can compute with the data available cover a limited age range of younger children and interrelatedness amongst scales tends to increase with age. We cannot carry out this comparison for the Denver-II and the BDI-2 screener since these correlations are not reported in the respective manuals.

4.3. Correlations with Other Variables

In a first exploration of test concurrent validity, we report the correlation of each of the test scales with a set of social background variables expected to be related with child development and that have been empirically shown so in a variety of contexts and countries, including the Latin American region. Specifically, we consider maternal education, the household wealth index, the quality of the home environment—as measured by the FCI scores for play materials and for play activity opportunities available in the home—prematurity and stunting.

The first panel in Table 4 shows that all scales in the Bayley-III are significantly correlated with the first four variables, except for gross motor which has very low significant correlations with maternal education, play materials and prematurity only. The correlations with prematurity and stunting are low for all scales and are not always significant. Specifically, the correlations between fine motor development and prematurity and between cognitive and

[17] Results using the 6-item version are robust to those using the 9-item version and available upon request.

gross motor development and stunting are not significant. These poor correlations may be explained by the relatively low prevalence of stunted and premature children in the sample.

Subsequent panels report the correlations for the short tests. As shown, most of the scales have low but significant correlations with at least two of the social background factors examined. The exceptions are all of the gross motor scales, as well as the expressive language scale in the SFI in children under 18 months and the personal-social scale in the Denver-II. The correlations between the scales in the short tests and prematurity and stunting are generally non-significant, except for the communication, motor and adaptive skills scales in the BDI-2. More generally, the BDI-2 shows the highest correlations among the short tests. However, these tend to be smaller than those observed for the Bayley-III.

Exploration of the correlations by age group (available upon request) shows a general tendency for these correlations to increase with age, particularly those between developmental outcomes and maternal education and household wealth. This is consistent with the socio-economic gradients of child development reported by many authors in the region and globally (Schady et al. 2015; Fernald et al. 2011; Hamadani et al. 2014; Rubio-Codina et al. 2015).

4.4. Concurrent Validity

4.4.1. Same Domain Scales, Entire Sample

Average correlations between the scales in the short tests and the Bayley-III for children across the entire age range (6-42 months) are reported in Table 5. Those between scales measuring the same developmental domain are highlighted in bold. Given that the Denver-II does not have a cognitive scale, we correlate its fine motor-adaptive scale with cognition. Also, with the exception of the SFI, the rest of the short tests combine receptive and expressive language into one communication/language scale. We correlate this scale with both the Bayley-III expressive and receptive language scales. Similarly, the BDI-2 motor scale combines fine and gross motor items and hence we correlate this scale both with the fine motor and gross motor scales in the Bayley-III. Furthermore, as discussed earlier, there is no matching Bayley-III scale for the personal-social or adaptive scales in the short tests available for analysis.

Overall, results show that the correlations between the Bayley-III and the shorter tests are low to moderate in magnitude. By domain, and of the multi-dimensional tests, the highest correlations are observed for expressive language, followed by gross motor. Correlations for expressive language are moderate: r =0.506 between the Bayley-III expressive language and the Denver-II language subscale, r =0.495 with BDI-2 communication, and r =0.395 with the ASQ-3 communication scale. For receptive language, the pattern of correlations follows the one observed for expressive language but correlations are generally 20% to 40% lower. Of the single-domain tests, concurrence is also higher for expressive rather than receptive language for the SFII but not for the SFI. Regarding gross motor development, the correlations between the Bayley-III gross motor and the matching scale in the short tests are particularly high for the WHO-Motor (r =0.703), moderate for the Denver-II (r =0.499), and low for the BDI-2 (r =0.339) and the ASQ-3 (r =0.325). Whilst significant, average concurrence for cognitive and fine motor development between matching scales is generally low for all short tests.

4.4.2. Same Domain Scales, by Age Groups

Table 6 reports concurrent validity by 12-month-of-age groups. The letters (a, b, c, etc.) next to some correlation values indicate if the correlation of interest is significantly larger (statistically speaking) than the correlation between another short test (reported in the footnote) and the Bayley-III. As explained in Section 3.2, we do this for each pair of correlations within an age group that exhibit large enough differences to merit testing for statistical significance. Figure 2 complements this analysis by graphically depicting the correlation between those scales in the Bayley-III and in the short tests measuring the same developmental domains by age. Observation of the patterns of concurrent validity by age suggests the study of cognitive, language, and fine motor development separate from the study of gross motor development. Therefore, we discuss them separately.

Cognitive, language, and fine motor. As shown in Table 6, the Denver-II and BDI-2 cognition/fine motor-adaptive, language/communication, and fine motor scales have similarly low but significant correlations with the corresponding Bayley-III scales at 6-18 months (r =[0.164, 0.315]). Concurrent validity increases slightly at 19-30 months (r =[0.256, 0.610]), even if it only reaches moderate or high levels for the language scales. Concurrence continues to improve over 30 months for all domains (r =[0.380, 0.702]), with language scales attaining once more the highest levels. For language, the correlations with the Bayley-III expressive language scale are always higher than those with the receptive language scale, across the age range.

A comparison across the multi-dimensional tests shows that the ASQ-3 scales consistently have lower correlations with the Bayley-III than the Denver-II and the BDI-2. These correlations are significantly lower in 16 of 24 comparisons (P <0.05). The only exception is the correlation between the ASQ-3 and the Bayley-III fine motor scales at 31-42 months, which is the same as that with the BDI-2. In the youngest group, the ASQ-3 correlations are generally trivial and non-significant for all domains. Moreover, the ASQ-3 problem solving scale does not significantly predict Bayley-III cognition until over 31 months. The highest correlations are for the ASQ-3 communication scale, which are low to moderate with the Bayley-III receptive language (r =[0.231, 0.402]), and moderate with the Bayley-III expressive language scale (r =[0.458, 0.560]) for children 19 months and over.

The SFI expressive language scale has slightly higher correlations with both Bayley-III language scales than the SFI receptive, although the differences are not statistically significant. It is possible that mothers find it easier to report words used by the child than words understood. In the youngest group, the SFI expressive scale has a low correlation with Bayley-III receptive language (r =0.373). Even so, this correlation is significantly larger than that with the Denver-II and BDI-2 language/communication scales (both P <0.05) and with the ASQ-3 communication (P <0.001). The correlation of the SFI expressive scale with the Bayley-III expressive language scale (r =0.242) is however similar to that with the other short tests. At 19-30 months, the SFII expressive has a low correlation with the Bayley-III receptive language (r =0.241), which is significantly lower than the Denver-II (P <0.05). Nevertheless, the SFII expressive has a high correlation with the Bayley-III expressive language scale (r =0.600), similar to the Denver-II and the BDI-2, and significantly higher than the ASQ-3 (P <0.05).

Gross motor. Gross motor scales behave differently from the other domains. As shown in Figure 2 and Table 6, the BDI-2 motor scale shows low correlations with the Bayley-III gross motor that change little throughout the age range (r =[0.311, 0.371]). However, the Denver-II and ASQ-3 gross motor correlations are moderate to high for children 6-18 months (*r* =[0.585, 0.654], significantly larger from the BDI-2, *P* <0.05) and then decrease for older children. While the Denver-II concurrence falls to moderate levels (*r* =[0.406, 0.426]), that of the ASQ-3 falls to low levels (*r* =[0.175, 0.218]), which is significantly lower than that for the Denver-II (*P* <0.05). The lower correlations with the Bayley-III gross motor scale for the BDI-2, even for the youngest children in the sample, may be due to the fact that the BDI-2 motor scale combines both fine and gross motor skills.

For children 6-15 months, the WHO-Motor has a high correlation with the Bayley-III gross motor (*r* =0.703). This correlation is higher than any other test for gross motor development, albeit it is only significantly higher from the BDI-2 (*P* <0.001).

4.4.3. Different Domains Scales

Occasionally, correlations between the Bayley-III and the short tests are higher between scales measuring different functions than those between scales measuring the same functions (Table 6). This happens less frequently as the children age. In the youngest group, the personal-social scales of the Denver-II and ASQ-3 correlate with cognition, language, and fine motor. For children over 18 months, the language scales are often significantly related to the Bayley-III cognition. Over 30 months, the Denver-II language scale correlation with cognition is significantly higher than the fine motor-adaptive scale (*P* <0.05). There are few other clear patterns in the cross-correlations.

The domain-specific tests also correlate with other domains. In the youngest group, the SFI expressive language shows significant but low correlations with cognition and fine motor, and the WHO-Motor shows low but significant correlations with cognition and expressive language.

4.4.4. Robustness

Before proceeding, we check robustness of these findings to the use of the Bayley-III composite scores—this is, to externally standardizing scores. The Bayley-III composite scores combine receptive and expressive language into one language scale, and fine and gross motor into one motor scale. Hence, we correlate the scales in the short tests to the three Bayley-III composite scores: cognitive, language, and motor. Results in Table A2 in Appendix III show that findings are generally similar to those reported above (Table 6), with only a few differences for children in the youngest age group. More specifically, for children under 19 months, the correlations between the gross motor scales in the short tests and the Bayley-III motor scale are no longer high, but attain moderate values. This is very likely due to the combination of the fine and gross motor scales in the Bayley-III composite motor score. For this age group, the correlation with Bayley-III cognition for the Denver-II fine motor-adaptive scale is also lower than that observed before (and trivial, i.e. *r* <0.20). For children 19 months and above, the pattern of correlations is very similar to the one found using the non-parametrically internally standardized scores.

Results are also qualitatively similar to a number of other robustness tests (available upon request): (i) standardizing raw scores internally within 2-months-of-age-intervals, each

interval with 25 to 51 children, following standard parametric procedures, (ii) using the original 6-item version of the ASQ-3, (iii) dividing the sample in 6-months-of-age groups instead of 12-months-of-age groups, (iv) controlling for prematurity before standardizing scores, (v) computing Spearman rank-order correlations, and (vi) computing canonical correlations. If anything, the results using both Spearman and canonical correlations strengthen our main findings.

4.4.5. Correlations by Household SES

One question that is often raised when assessing the development of young children in the context of large impact evaluations of programs targeted to poor and vulnerable families is whether the low levels of maternal education, often present in these populations, will affect the quality of the data collected. More educated mothers might be more knowledgeable about developmental milestones, pay more attention to observing their child's abilities, or report them more accurately. Hence, the ability of a test to measure child development may decrease with the caregiver ability to report on the child´s abilities, and more so the more items that are obtained by caregiver report.

This study provides a good setting to empirically investigate the extent to which the relative ability of the short tests to measure child development varies with caregiver characteristics. We consider household wealth, in turn associated with caregiver (parental) education (r =0.47), to be a good proxy of the caregiver's ability to accurately report on the child's developmental level, and replicate the analysis of concurrent validity for households in the bottom and top 25% of the household wealth distribution in the sample.[18] We test whether the size of the correlation is significantly different between the 25% poorest and 25% richest households using bootstrapping methods. To ensure sufficient power, we restrict this exercise to the complete sample of children (6-42 months).

Table 7, which is analogous to Table 5, reports the concurrent validity by domain for children 6-42 months in the bottom 25% of the household wealth distribution (left set of columns) and in the top 25% (right set of columns). A comparison of the Pearson correlation coefficients between the two types of households does not seem to reveal any clear pattern nor marked differences in correlation sizes: the difference in correlation coefficients ranges from 0.010 to 0.230 in absolute values, with coefficients being sometimes larger for the richest in the sample and others smaller. Note that, given the findings so far, we are restricting this exercise to the comparison of correlations between matching developmental domains. This gives us a total of 18 possible comparisons. For each comparison, the formal test of significance always fails to reject the null of no significant differences, which indicates no differential reporting of caregivers by SES status.

4.5. Tests Training and Administration Costs

Table 1, already presented in Section 2.2, reports the costs of training and administering the tests, including the cost of purchase, per child unitary fee, and time and difficulty to train and administer. As discussed earlier, the assessments on the difficulty to train and to administer the test are by trainer report and hence subject to a certain degree of subjectivity. Similarly, the number of days to train is based on our experience in this and other studies but may vary

[18] We use the household wealth index, as opposed to maternal education, for this analysis since the distribution has more variability, is more continuous, and is normally distributed.

as a function of the qualifications and previous experience of the trainee and the trainer. Time to administer the test *in the study* is the average time recorded by the trainer for those administrations that were supervised during data collection.

Overall, Table 1 shows that the costs of the kit of test materials and the per child administration fee (unitary cost of record form) are substantially higher in the Bayley-III than in any of the short tests. The Bayley-III also takes longer to train and to administer (83 minutes, on average), and requires the most skill to learn and give. Of the multi-dimensional short tests, the BDI-2 takes longest (59 minutes, on average) and is the most expensive. The Denver-II and ASQ-3 are intermediate both in terms of time and cost, with the ASQ-3 being 7-8 minutes shorter than the Denver-II on average, and having a slightly more expensive kit but photocopiable record forms.[19] As expected, the single-domain tests are the shortest and cheapest of the short tests, both taking less than 10 minutes to give and the WHO-Motor being free.

Training time also increases with the length of the test and with the number of items that are scored by direct administration or by observation of the child's abilities during the assessment. Those scales or tests relying mostly on caregiver reports are the easiest to train and to administer, as reported by the trainer. The easiest of all are the vocabulary check-lists in the SFs. And a curios note: when asked about the tests they enjoyed the most, caregivers responded the WHO-Motor and the Bayley-III, which are those that require the least contribution (almost none) from the caregiver. The scales we administered from these tests were entirely collected by direct child assessment.

5. Conclusions: Choice of Test and Final Remarks

We have examined the use of three multi-dimensional screeners—the ASQ-3, the Denver-II and the BDI-2 (Battelle screener)—and two single-domain tests of child development—the SFs (Mac-Arthur Bates CDIs) and the WHO-Motor—in a sample of children 6-42 months representative of low and lower-middle-income households in the city of Bogota, Colombia, specifically focusing on their reliability, validity and feasibility when used in situation mimicking an at-scale evaluation. Many of these tests have been previously used in large scale studies in LMICs (Fernald et al. 2012; Macours, Schady, and Vakis 2012; Fernald and Hidrobo 2011).

Throughout the analysis, we have considered the Bayley-III as the gold standard. Even if this test has not been standardized in Colombia, we have shown that it is valid in our sample and therefore appropriate as a gold standard. We have shown it has good internal and test-retest reliability, and that it relates to socio-economic child, maternal, and household characteristics as theoretically expected. In fact, in an earlier study we had reported that the test scores showed differences by wealth quartiles from the first year of life that increased to 42 months in this sample (Rubio-Codina et al. 2015). The scales also showed acceptable levels of predictive validity with measures of cognition, language, and school readiness at age 5 in a contemporaneous Colombian study by the same researchers (personal communication). Furthermore, the correlations of the Bayley-III scales among each other are also similar to those reported in the test manual (Bayley 2006).

[19] The Denver-II has recently been discontinued. Nonetheless, test materials (manual and report forms) are still available for download from the publisher's website in English and Spanish.

Reliability of the short tests analyzed is generally acceptable. Although sample sizes are very small for the computation of test-retest reliabilities, values are good for all short tests, except for the fine motor scale in the ASQ-3 and the gross motor and personal-social subscales in the Denver-II. Similarly the internal consistency reliability shows good or acceptable values, with the exception of the ASQ-3 and some scales in the BDI-2, particularly for children in the middle age group. Internal consistency does not necessarily improve with age. The scales in the short tests are also shown to be correlated amongst each other as expected, achieving values similar to those reported in the tests manuals for those tests for which they are available. This indicates congruence between scales and a level of interrelatedness between developmental domains.

Regarding test validity, all short tests correlate with a set of standard child, maternal and household socio-economic variables as expected, although correlations are much lower than those observed for the Bayley-III. The pattern of concurrent validity with the Bayley-III varies by age and domain, particularly for the multi-dimensional tests, which also cover the entire age range under study. Generally speaking, concurrent validity increases with age for the language, cognitive, and fine motor scales. Concurrence of these scales in the Denver-II and BDI-2 is low but significant below 19 months, moderate at 19-30 months, and moderate-to-high over 30 months, with language usually showing the highest levels of concurrent validity over 19 months. Throughout the age range, the ASQ-3 has consistently poorer concurrence in these scales than the Denver-II and the BDI-2. In fact, it seems to be non-informative for children younger than 19 months. Findings on the ASQ-3 hold if we use the 6-item version of the test. With the exception of the BDI-2, the gross motor scales behave differently, having high concurrent validity below 19 months, which then declines. Of the single-domain tests, the WHO-Motor has high concurrence with the Bayley-III gross motor scale up to age 15 month; and the expressive language in the SFII shows high correlations between 19-30 months. These findings hold to standardizing the test scores using different methods and to the use of rank correlations, amongst other robustness tests.

Examination of concurrent validity by household SES shows no statistically significant differences between the 25% poorest and the 25% richest households in the sample. This seems to dismiss any concerns related to a higher ability to report on child development by more educated caregivers, which could have affected the performance of some tests, particularly those with a higher content of items administered by caregiver report.

5.1. Choice of Test

The choice of tests depends on the availability of time, funds and qualified testers, all of which are usually limited in large surveys. The choice also depends on test validity, the amount of adaptation required, age of the children and study objectives. The main outcomes of interest, for example, may vary depending on whether the aim is to establish the broad developmental profile of a population or to evaluate an intervention, as well as on the type of intervention being evaluated.

All the multi-dimensional tests spread over the entire age range and the concurrent validity of the cognitive, language, and fine motor scales found is little different between the BDI-2 and the Denver-II. However, the BDI-2 is much longer to administer and more expensive (both the cost of the test kit and the unitary cost of the record form). It also requires more adaptations and translations, as well as a longer training. The Denver-II and the ASQ-3

require similar amounts of materials for administration, and they both take less time to give than the BDI-2. Nonetheless, the lower concurrent validity of the ASQ-3 for all scales, suggests that the Denver-II is the most suitable—reliable, valid and feasible—for use at-scale. In fact, the poor validity of the ASQ-3 below 30 months is concerning given the test is increasingly used in large scale studies (Fernald et al. 2012; Martinez and Naudeau 2012). While it is possible that the language modifications to Spanish may have changed the psychometric properties of the ASQ-3, these findings do not concern its validity as a screener of high risk children. In further support of the Denver-II, note that the test, administered in the home, was sensitive to the impact of a cash transfer program in Nicaragua (Macours, Schady, and Vakis 2012).

The single-domain tests are the most feasible to give, being short, inexpensive, and requiring little training. Whilst, their age range is limited, they offered reasonable levels of concurrence for the domains and ages for which they are available. Therefore, they might be of consideration for survey work. The WHO-Motor has shown to be highly valid for gross motor development under 16 months. In addition, it has low correlations with cognition and expressive language. This concurs with findings from the Bangladeshi study discussed earlier (Hamadani et al. 2013). Note, however, that monthly assessments were used in Bangladesh and may be more accurate than one examination only, as in the present study.

Similarly, the SFI and SFII expressive have at least as good validity as the language scales of the multi-dimensional tests and low correlations with cognition and fine motor under 19 months in Bogota. In Bangladesh, vocabulary reports locally developed from the SFII also had moderate concurrent validity (r=0.41, P<0.01) with the Bayley-II MDI at 18 months, and predicted IQ at 5 years (r=0.37, P <0.01) (Hamadani et al. 2010). An advantage of maternal reports of early vocabulary is that disadvantaged young children, who tend to be inhibited in LMICs, do not have to speak to the tester. A disadvantage is that a new inventory has to be 'developed' for every new language, which is time consuming and requires skill.[20] In addition, some adaptations may be needed when using the same language in different countries/contexts (dialectal adaptations). Nonetheless, this is feasible and in fact the SFs are already available in many languages.

There is general agreement that multi-dimensional tests are most desirable (Fernandes et al. 2014). But, where resources are limited, it may be possible to use selected scales of a test or single-domain tests, depending on the children's age and purpose of the assessment. For example, to evaluate psychosocial stimulation programs that rarely target gross motor development, the language and fine motor-adaptive Denver-II scales could be used. For nutritional interventions, however, which more often affect motor development in younger children, the WHO-Motor would be useful in children under 16 months; especially since it has low but significant correlations with the cognitive and expressive language scales. If language is the focus of interest and children are under 30 months, then the SFI and SFII could be used without the receptive scale.

Overall, the low-to-moderate concurrent validity of all tests except the gross motor scales in the youngest children concurs with reported difficulties in assessing young children's development, particularly at large scale (Frongillo et al. 2014; Fernald et al. 2009; Fernandes et al. 2014). As a result, with the exception of the gross motor domain, all tests seem to have

[20] http://mb-cdi.stanford.edu/adaptations.html. The CDI Advisory Board at Stanford provides guidance and support to these efforts, whenever possible.

limited validity under 18 months. Nonetheless, the findings in this study need to be complemented with the examination of predictive validity to be certain since concurrent and predictive validity may not necessarily be closely related. For example, the Bangladeshi language test had moderate concurrent validity with the Bayley-II at 18 months but similar predictive validity of later IQ (Jena D. Hamadani et al. 2010). Similarly, more research on developing or modifying tests for children under 24 months would be desirable.

5.2. Concluding Remarks

Measuring ECD outcomes for very young children at-scale is challenging. However, multi-dimensional screeners and single-domain tests offer feasible, reliable alternatives. Concurrent validity varies by domain and age. The scales with the highest concurrent validity are gross motor under 19 months and language above. The Denver-II is the most feasible and valid multi-dimensional test and the ASQ-3 generally shows poor performance under 31 months. Investigation of predictive validity and sensitivity to interventions is needed to further support these findings, which should be helpful in the selection of instruments and design of future large scale studies interested in the measurement of child development. We are currently preparing a follow-up study to examine the relative predictive validity of IQ, language development, executive functioning and school achievement, of the short tests and the Bayley-III at 6-9 years-of-age. Results from this follow-up study will complement the findings reported herein.

References

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

Attanasio, O. P. 2015. “The Determinants of Human Capital Formation during the Early Years of Life: Theory, Measurement, and Policies.” *Journal of the European Economic Association* 13 (6): 949–97. doi:DOI: 10.1111/jeea.12159.

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

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

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

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

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

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

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

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

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

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

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

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

Frankenburg, W. K., J. Dodds, P. Archer, B. Bresnick, P. Maschka, N. Edelmann, and H. Shapiro. 1990. *The DENVER II Technical Manual*. Denver, CO: Denver Developmental Materials.

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

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

Frongillo, E. A., S. M. Sywulka, and P. Kariger. 2003. "UNICEF Psychosocial Care Indicators Project."

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

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

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

Hamadani, J. D., F. Tofail, S. N. Huda, D. S. Alam, D. a. Ridout, O. Attanasio, and S. M. Grantham-McGregor. 2014. "Cognitive Deficit and Poverty in the First 5 Years of Childhood in Bangladesh." *Pediatrics* 134 (4): e1001–8. doi:10.1542/peds.2014-0694.

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

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

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

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

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

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

Jackson-Maldonado, D., D. Thal, V. Marchman, T. Newton, L. Fenson, and B. Conboy. 2003. *Mac Arthur Inventarios Del Desarrollo de Habilidades Comunicativas. User's*

Guide and Technical Manual. Baltimore, MD: Paul H. Brookes Publishing Co.

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

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

Martinez, S., and S.Naudeau. 2012. “The Promise of Preschool in Africa : A Randomized Impact Evaluation of Early Childhood Development in Rural Mozambique.”mimeo The World Bank, Washington D.C.

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

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

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

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

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

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

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

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

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

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

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

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

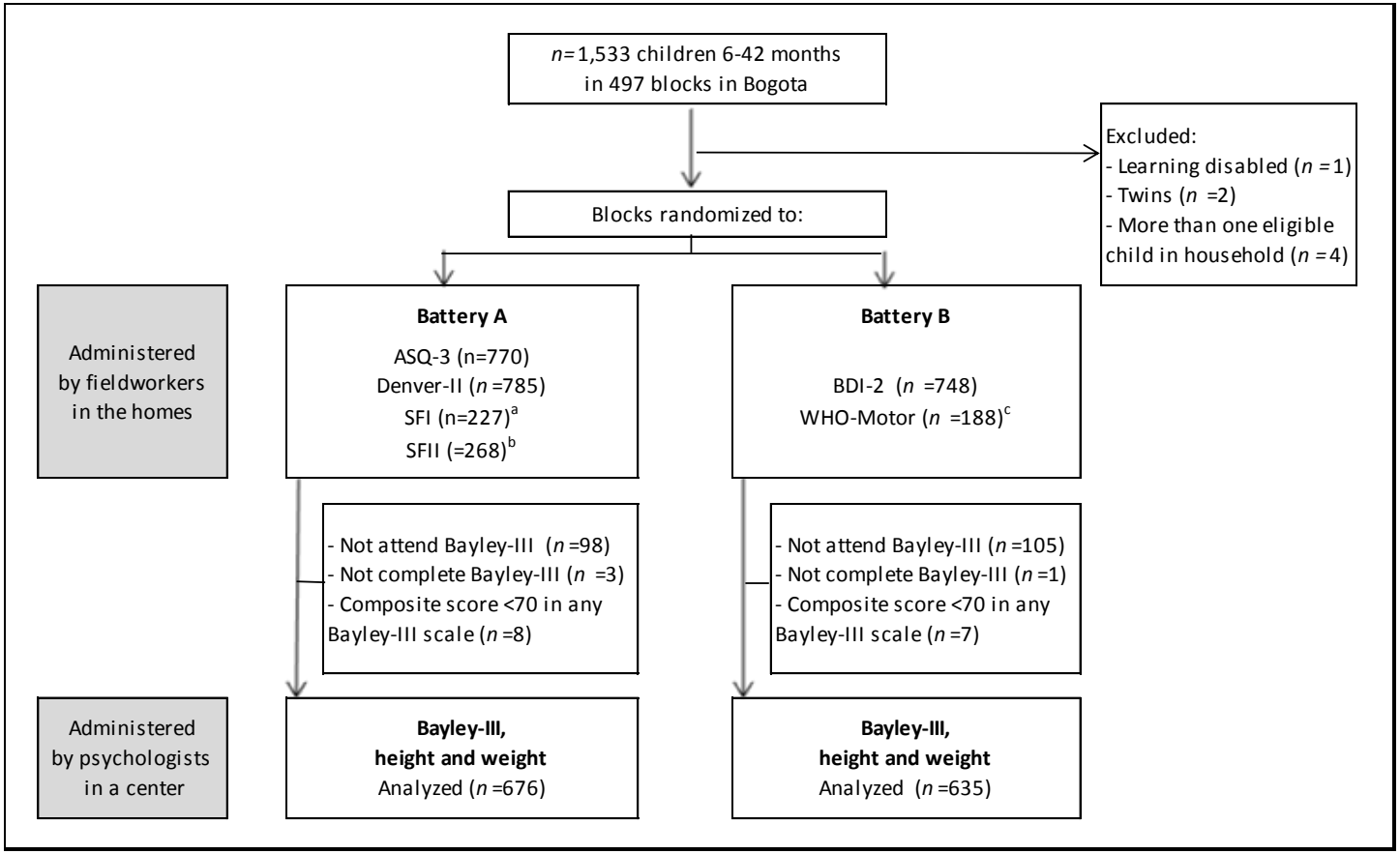
World Wealth Organization. 1983. “Measuring Change in Nutritional Status. Guidelines for Assessing the Nutritional Impact of Supplementatry Feeding Programmes for Vulnerable Groups.” Geneva.

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

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

Tables and Figures

Figure 1: Flow Diagram of Study Participants



[a] Children 8-18 months.

[b] Children 19-30 months.

[c] Children 6-15 months.

Figure 2: Concurrent Validity between the Bayley-III and Short Tests by Matching Domain and Age Group

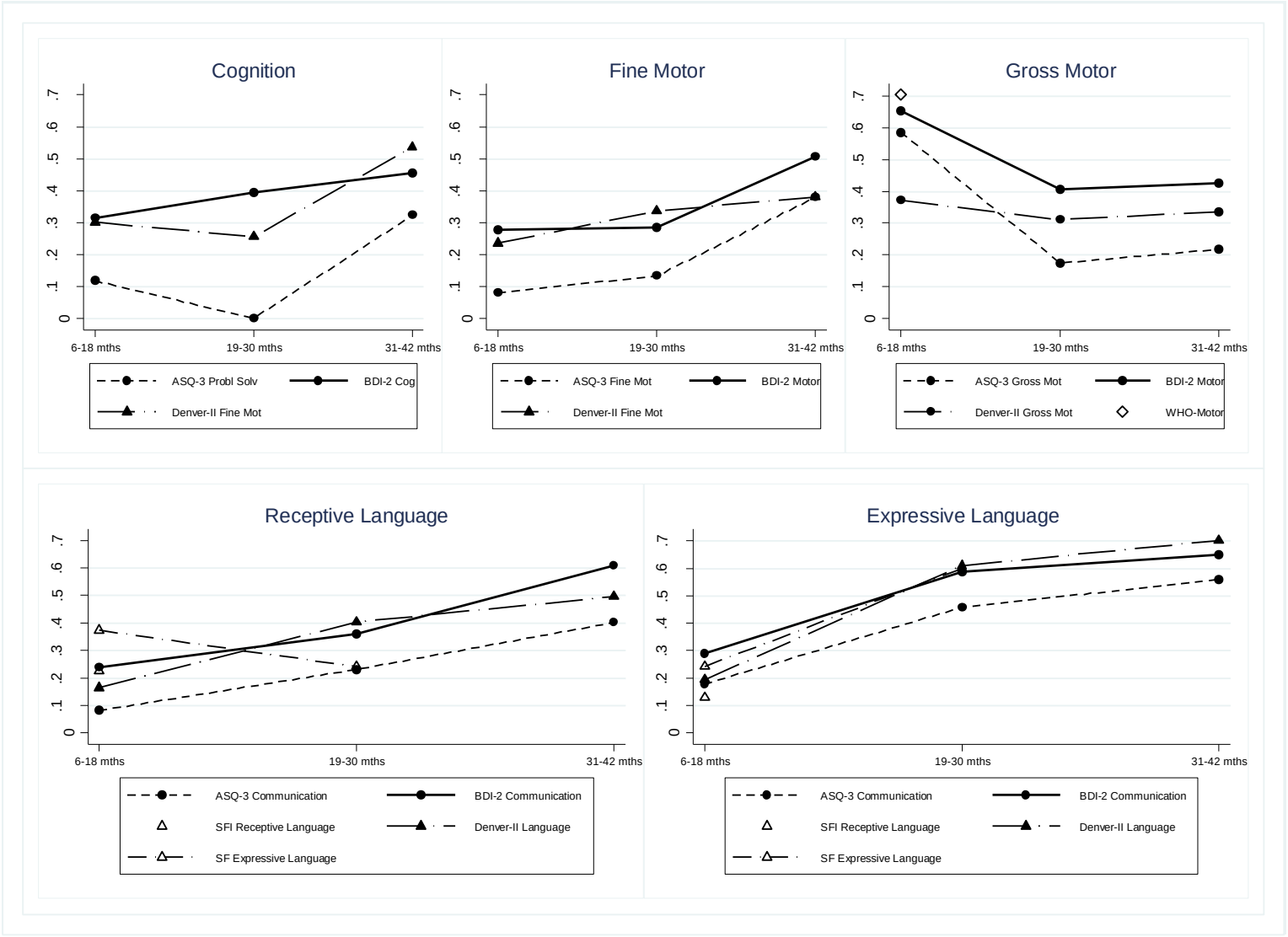


Table 1: Characteristics of the Bayley-III and the Short Tests

Test	Scales Included in Study and Number of Items[a]	Age Range Test (months)	Age Range Study Children (months)	Cost (USD)[b]	Minutes to Administer[c]	Minutes to Administer in Study[d]	Days to Train[e]	Difficulty to Train[f]	Difficulty to Administer[f]
Bayley-III	Cognitive (91) [21] Receptive Language (48) [18] Expressive Language (49) [16] Fine Motor (66) [18] Gross Motor (72) [16]	0-42	6-42	$1,050 kit + $9.34 pc	30-95	(*n* =36) 83.2 (18.8)	15 + practice	high	high
ASQ-3	Problem Solving (6) [6.5] Communication (6) [6.5] Fine Motor (6) [6.4] Gross Motor (6) [6.6] Personal-Social (6) [6.4]	1-66	6-42	$275 kit $295 materials	10-15	(*n* =32) 19.7 (8.2)	6 + practice	medium	medium
Denver-II	Fine Motor-Adaptive (29) [9] Language (39) [10] Gross Motor (32) [9] Personal-Social (25) [9]	0-71	6-42	$200 kit + $0.45 pc	15-20	(n =32) 27 (10.5)	7 + practice	high	medium/ high
SFI (MacArthur)	Receptive Language (104) [104] Expressive Language (104) [104]	8-18	8-18	$90 kit + $1 pc	10	(*n* =8) 8.6 (1.9)	0.5 + minimal practice	low	low
SFII (MacArthur)	Expressive Language (100) [100]	19-30	19-30		10	(*n* =10) 8.2 (3.3)			
BDI-2 (Battelle)	Cognitive (20) [9] Communication (20) [9] Motor (20) [9] Personal-Social (20) [10] Adaptive Skills (20) [9]	0-83	6-42	$405.70 kit + $3.08 pc	10-30	(*n* =30) 59 (13.0)	8 + practice	high	high
WHO-Motor	Gross motor (6) [6]	4-24	6-15	free	10	(*n* =9) 6 (2.7)	1 + practice	medium	medium

[a] Number of total items in the scale in parentheses, and average number of items assessed per study participant in brackets.

[b] Information on costs last consulted on line in March 2016 from the publishers website (see details in Appendix I). pc is 'per child' administration fee. Test kits include record forms in packages of 100 for the Denver-II, packages of 30 for the BDI-2, and packages of 25 for the rest of the short tests and the Bayley-III. The WHO-Motor was not available in Spanish; and only parts of the BDI-2 were available. The Bayley-III is available in Spanish since mid 2015. All other tests and manuals were available in Spanish.

[c] Time to administer the test as reported in the publishers' website.

[d] Data are Mean (Standard Deviation) in minutes, as recorded by the trainer during supervision activities.

[e] Number of days are estimates on the basis of our experience and are subject to change depending on the trainee and trainer qualifications and experience, amongst others.

[f] As reported by the trainer.

Table 2: Characteristics of Children in the Study Sample and their Families by Battery

	Battery A (n_A=676)	Battery B (n_B=635)	p-value difference in batteries
I. Child Characteristics			
Child's age, %			
6-18 months	33.7	33.9	0.960
19-30 months	33.6	35.7	0.410
31-42 months	32.7	30.4	0.371
Girls, %	47.6	51.0	0.220
Premature (gestional age <37 weeks), %	15.2	15.1	0.952
Birth weight[a], gr, mean (SD)	3066 (514)	3015 (510)	0.087
Stunted (z score height-for-age <-2SD)	16.9	17.7	0.710
II. Parental Characteristics			
Mother's age[a], y, mean (SD)	27.2 (6.9)	26.6 (6.4)	0.106
Mother's education[a], y, mean (SD)	10.3 (3.4)	10.4 (3.3)	0.541
Mother has more than secondary education	31.0	30.1	0.729
Mother works (paid or unpaid employment)	49.7	53.9	0.134
Mother gave birth before age 18	13.3	13.5	0.903
Father's education[a], y, mean (SD)	8.1 (4.0)	8.7 (4.0)	0.026
Father has more than secondary education	29.1	27.6	0.629
Father absent (deceased or away)	32.0	33.5	0.540
III. Household Characteristics			
Socio-economic sector (strata), %			
1. Lowest	30.3	29.8	0.825
2.	32.7	37.0	0.101
3.	36.1	32.6	0.183
4. Highest	0.9	0.6	0.592
Household wealth index	-0.089 (1)	0.05 (1)	0.013
Number of varieties of play materials	4.8 (2.3)	4.8 (2.4)	0.682
Number of varieties of play activities	3.7 (1.9)	3.6 (1.8)	0.235
IV. Bayley-III Raw Scores, mean (SD)			
Cognitive	58.7 (14.5)	58.8 (13.5)	0.876
Receptive language	25.3 (9.3)	25.6 (8.9)	0.528
Expressive language	25.0 (10.3)	25.5 (10.2)	0.352
Fine motor	39.5 (10.1)	39.1 (10.0)	0.442
Gross motor	52.2 (11.7)	52.6 (11.4)	0.461

[a]Missing data for: birth weight (n_A=638, n_B=552), mother's age and mother's employment status (n_A=668, n_B=618), mother's education (n_A=674, n_B=633), father's education (n_A=639, n_B=576). SD is Standard Deviation. Household wealth index, and number of varieties of play materials and of play activities are computed as described in the text.

Table 3: Test-retest Reliability and Internal Consistency (Overall and by Age Groups) of the Bayley-III and Short Tests

	Test-retest ICC	Cronbach alpha 6-42 mths	Cronbach alpha 6-18 mths	Cronbach alpha 19-30 mths	Cronbach alpha 31-42 mths
Bayley-III Raw Scores	(*n* = 20)	(*n* = 1311)	(*n* = 443)	(*n* = 454)	(*n* = 414)
Cognition	0.96	0.97	0.94	0.90	0.82
Receptive Language	0.96	0.96	0.85	0.90	0.79
Expressive Language	0.98	0.96	0.88	0.90	0.91
Fine Motor	0.98	0.96	0.92	0.85	0.85
Gross Motor	0.98	0.97	0.96	0.85	0.78
Bayley-III Composite Scores					
Cognition	0.96	0.97	0.94	0.90	0.82
Language	0.97	0.98	0.93	0.94	0.92
Motor	0.98	0.98	0.97	0.91	0.88
ASQ-3 (9 items)	(*n* = 12)	(*n* = 664)	(*n* = 221)	(*n* = 224)	(*n* = 219)
Problem Solving	0.80	0.60	0.54	0.62	0.66
Communication	0.92	0.68	0.58	0.71	0.78
Fine Motor	0.37	0.57	0.63	0.44	0.65
Gross Motor	0.90	0.70	0.76	0.66	0.68
Personal-Social	0.73	0.55	0.57	0.44	0.65
ASQ-3 (6 items)	(*n* = 12)	(*n* = 664)	(*n* = 221)	(*n* = 224)	(*n* = 219)
Problem Solving	0.80	0.42	0.45	0.32	0.51
Communication	0.92	0.52	0.47	0.55	0.55
Fine Motor	0.37	0.44	0.49	0.37	0.45
Gross Motor	0.90	0.55	0.72	0.47	0.39
Personal-Social	0.73	0.38	0.41	0.33	0.40
Denver-II	(*n* = 12)	(*n* = 658)	(*n* = 225)	(*n* = 221)	(*n* = 212)
Language	0.93	0.93	0.85	0.85	0.90
Fine Motor-Adaptive	0.83	0.91	0.86	0.81	0.78
Gross Motor	0.53	0.90	0.90	0.78	0.74
Personal-Social	0.49	0.91	0.90	0.76	0.76
SFI (MacArthur)	(*n* = 12)	(*n* = 192)	(*n* = 192)		
Receptive Language	0.99	0.97	0.97		
Expressive Language		0.92	0.92		
SFII (MacArthur)		(*n* = 226)		(*n* = 226)	
Expressive Language	NA	0.98		0.98	
BDI-2 (Battelle)	(*n* = 11)	(*n* = 635)	(*n* = 215)	(*n* = 227)	(*n* = 193)
Cognitive	0.92	0.79	0.62	0.40	0.72
Communication	0.94	0.89	0.76	0.67	0.78
Motor	0.98	0.88	0.83	0.63	0.54
Personal-Social	0.71	0.84	0.73	0.65	0.76
Adaptive Skills	0.90	0.84	0.71	0.61	0.62
WHO-Motor	(*n* = 11)	(*n* = 152)	(*n* = 152)		
Gross Motor	0.80	0.86	0.86		

Table 4: Correlations of the Bayley-III and the Short Tests with Socio-economic Variables

	Maternal Education	Wealth Index	Play Activities	Play Materials	Prematurity	Stunting
Bayley-III (*n* =1311)						
Cognition	0.210***	0.235***	0.189***	0.271***	-0.096***	-0.051
Receptive Language	0.216***	0.191***	0.214***	0.248***	-0.056*	-0.080**
Expressive Language	0.206***	0.224***	0.209***	0.243***	-0.063*	-0.069*
Fine Motor	0.124***	0.145***	0.119***	0.179***	-0.038	-0.082**
Gross Motor	0.079**	0.034	0.023	0.056*	-0.092***	-0.051
ASQ-3 (*n* =664)						
Problem Solving	0.127**	0.071	0.176***	0.177***	-0.008	0.002
Communication	0.142***	0.136***	0.222***	0.156***	-0.012	-0.009
Fine Motor	0.063	0.067	0.167***	0.133***	0.012	-0.011
Gross Motor	-0.025	0.046	0.069	0.019	0.005	-0.012
Personal-Social	0.019	0.034	0.152***	0.088*	0.007	0.062
Denver-II (*n* =658)						
Language	0.170***	0.165***	0.184***	0.173***	-0.012	-0.049
Fine Motor-Adaptive	0.102**	0.121**	0.097*	0.109**	-0.063	-0.027
Gross Motor	0.022	0.020	-0.021	-0.011	0.018	-0.068
Personal-Social	-0.034	-0.019	0.064	0.010	0.022	0.006
SFI (MacArthur) (*n* =192)[a]						
Receptive Language	0.147*	0.127	0.267***	0.251***	-0.049	-0.012
Expressive Language	0.040	-0.060	-0.007	-0.005	-0.092	0.024
SFII (MacArthur) (n =226)[b]						
Expressive Language	0.136*	0.094	0.229***	0.200**	-0.058	0.021
BDI-2 (Battelle) (*n* =635)						
Cognitive	0.202***	0.173***	0.164***	0.181***	0.004	-0.056
Communication	0.210***	0.176***	0.224***	0.245***	-0.080	-0.152***
Motor	0.139***	0.163***	0.135***	0.179***	-0.016	-0.102*
Personal-Social	0.144***	0.136***	0.240***	0.231***	-0.012	-0.057
Adaptive Skills	0.074	0.094*	0.276***	0.193***	-0.029	-0.123**
WHO-Motor (*n* =152)[c]						
Gross Motor	-0.036	0.008	0.082	0.018	-0.103	0.116

* p<0.05, ** p<0.01, *** p<0.001. [a]Children 8-18 months. [b]Children 8-30 months. [c]Children 6-15 months.

Table 5: Correlation among Bayley-III Scales and between Scales in the Bayley-III and the Short Tests, Children 6-42 Months

	Bayley-III, 6-42 months				
	Cognitive	**Receptive Language**	**Expressive Language**	**Fine Motor**	**Gross Motor**
Bayley-III			*n* = 1311		
Cognitive	1				
Receptive Language	0.544***	1			
Expressive Language	0.483***	0.563***	1		
Fine Motor	0.529***	0.461***	0.413***	1	
Gross Motor	0.369***	0.356***	0.306***	0.380***	1
ASQ-3			*n* = 664		
Problem Solving	**0.146*** **	0.156***	0.221***	0.151***	0.048
Communication	0.199***	**0.236*** **	**0.395*** **	0.164***	0.126**
Fine Motor	0.172***	0.157***	0.192***	**0.200*** **	0.163***
Gross Motor	0.066	0.073	0.043	0.067	**0.325*** **
Personal-Social	0.100*	0.134***	0.172***	0.124**	0.098*
Denver-II			*n* = 658		
Language	0.329***	**0.401*** **	**0.506*** **	0.246***	0.193***
Fine Motor-Adaptive	**0.386*** **	0.329***	0.308***	**0.354*** **	0.210***
Gross Motor	0.216***	0.234***	0.183***	0.204***	**0.499*** **
Personal-Social	0.244***	0.215***	0.226***	0.195***	0.184***
SFI & SFII (MacArthur)			*n* = 418		
Receptive Language[a]	0.187**	**0.224** **	**0.130**	0.088	0.168*
Expressive Language[b]	0.206***	**0.299*** **	**0.441*** **	0.131**	0.152**
BDI-2 (Battelle)			*n* = 635		
Cognitive	**0.363*** **	0.319***	0.337***	0.308***	0.210***
Communication	0.343***	**0.349*** **	**0.495*** **	0.237***	0.263***
Motor	0.269***	0.220***	0.252***	**0.316*** **	**0.339*** **
Personal-Social	0.124**	0.161***	0.209***	0.040	0.058
Adaptive Skills	0.153***	0.225***	0.233***	0.190***	0.208***
WHO-Motor			*n* = 152[c]		
Gross Motor	0.224**	0.126	0.282***	0.061	**0.703*** **

* p<0.05, ** p<0.01, *** p<0.001. Pearson correlations on internally standardised scores, Standard Errors (SE) computed using bootstrap stratifying by age category and socio-economic sector (*n* =1000 replications). Matching scales bolded. [a]Children 8-18 months; [b]Children 8-30 months; [c]Children 6-15 months.

Table 6: Correlation among Bayley-III Scales and between Scales in the Bayley-III and the Short Tests, by Age Group

	Bayley-III, 6-18 months					Bayley-III, 19-30 months					Bayley-III, 31-42 months				
	Cognitive	Receptive Language	Expressive Language	Fine Motor	Gross Motor	Cognitive	Receptive Language	Expressive Language	Fine Motor	Gross Motor	Cognitive	Receptive Language	Expressive Language	Fine Motor	Gross Motor
Bayley-III			*n* = 443					*n* = 454					*n* = 414		
Cognitive	1					1					1				
Receptive Language	0.437***	1				0.604***	1				0.590***	1			
Expressive Language	0.356***	0.502***	1			0.494***	0.554***	1			0.603***	0.639***	1		
Fine Motor	0.533***	0.490***	0.408***	1		0.525***	0.435***	0.377***	1		0.528***	0.458***	0.457***	1	
Gross Motor	0.333***	0.329***	0.232***	0.354***	1	0.392***	0.407***	0.350***	0.421***	1	0.381***	0.329***	0.334***	0.363***	1
ASQ-3			*n* = 221					*n* = 224					*n* = 219		
Problem Solving	**0.119**	0.010	0.071	0.062	0.026	**0.001**	0.075	0.133*	0.091	0.005	**0.323*** **	0.374***	0.454***	0.292***	0.111
Communication	0.104	**0.082**	**0.178****	0.142*	0.164*	0.141*	**0.231*** **	**0.458*** **	0.069	0.067	0.361***	**0.402*** **	**0.560*** **	0.286***	0.147*
Fine Motor	0.084	0.084	0.077	**0.082**	0.151*	0.135*	0.131*	0.171**	**0.134***	0.176**	0.297***	0.256***	0.331***	**0.380*** **	0.164*
Gross Motor	0.147*	0.148*	0.047	0.105	**0.585*** ** [b]	-0.053	0.013	0.025	0.010	**0.175****	0.110	0.060	0.059	0.090	**0.218*** **
Personal-Social	0.208**	0.184**	0.166*	0.208**	0.070	0.033	0.082	0.121	0.065	0.109	0.063	0.140*	0.240***	0.102	0.119
Denver-II			*n* = 225					*n* = 221					*n* = 212		
Language	0.206**	**0.238*** ** [a]	**0.290*** **	0.187*	0.125	0.224***	**0.361*** ** [a,d]	**0.587*** ** [a]	0.111	0.175**	0.560*** [e]	**0.608*** ** [a]	**0.650*** **	0.443***	0.284***
Fine Motor-Adaptive	**0.315*** ** [a]	0.279***	0.168*	**0.277*** ** [a]	0.153*	**0.395*** ** [a]	0.339***	0.345***	**0.286*** ** [a]	0.257***	**0.455*** **	0.377***	0.426***	**0.507*** ** [a]	0.229***
Gross Motor	0.264***	0.270***	0.085	0.180**	**0.654*** ** [b]	0.133*	0.171**	0.207***	0.198**	**0.406*** ** [a]	0.256***	0.263***	0.270***	0.239***	**0.426*** ** [a]
Personal-Social	0.366***	0.279***	0.296***	0.240***	0.244***	0.099	0.174**	0.185**	0.182**	0.197**	0.274***	0.194**	0.200**	0.166*	0.111
SFI & SFII (MacArthur)			*n* = 192[f]					*n* = 226							
Receptive Language	0.187**	**0.224*** **	**0.130**	0.088	0.168*										
Expressive Language	0.258***	**0.373*** ** [a, b, c]	**0.242*** **	0.204***	0.176*	0.168*	**0.241*** **	**0.600*** ** [a]	0.077	0.134*					
BDI-2 (Battelle)			*n* = 215					*n* = 227					*n* = 193		
Cognitive	**0.302*** ** [a]	0.223***	0.209**	0.274***	0.244***	**0.256*** ** [a]	0.297***	0.327***	0.253***	0.146	**0.536*** ** [a]	0.444***	0.484***	0.404***	0.243***
Communication	0.205**	**0.164***	**0.194****	0.195**	0.229***	0.350***	**0.403*** ** [a]	**0.610*** ** [a]	0.196**	0.245***	0.488***	**0.496*** **	**0.702*** ** [a]	0.329***	0.325***
Motor	0.147*	0.161*	0.109	**0.236*** **	**0.371*** **	0.288***	0.231***	0.268***	**0.337*** ** [a]	**0.311*** **	0.379***	0.273**	0.386***	**0.380*** **	**0.335*** **
Personal-Social	0.025	0.072	0.057	0.001	0.077	0.145	0.193**	0.286***	0.019	-0.015	0.210**	0.226**	0.293***	0.110	0.126
Adaptive Skills	0.090	0.206**	0.228***	0.137*	0.271***	0.168*	0.218***	0.234***	0.241***	0.257***	0.201**	0.255***	0.240***	0.189*	0.077
WHO-Motor			*n* = 152[g]												
Gross Motor	0.224**	0.126	0.282***	0.061	**0.703*** ** [b]										

* p<0.05, ** p<0.01, *** p<0.001. Pearson correlations on internally standardised scores, Standard Errors (SE) computed using bootstrap stratifying by age category and socio-economic sector (*n* =1000 replications). Matching scales bolded.

[a] Concurrence larger than ASQ concurrence, matching domain; [b] Concurrence larger than BDI-2 concurrence, matching domain; [c] Concurrence larger than Denver-II concurrence, matching domain; [d] Concurrence larger than SFII concurrence, matching domain; [e] Concurrence larger than Denver-II fine motor-adaptive concurrence with Bayley-III cognitive. [f] Children 8-18 months; [g] Children 6-15 months.

Table 7: Correlation between Scales in the Bayley-III and the Short Tests for the 25% Poorest and 25% Richest in the Sample, Children 6-42 Months

	Bayley-III, 25% Poorest					Bayley-III, 25% Richest				
	Cognitive	**Receptive Language**	**Expressive Language**	**Fine Motor**	**Gross Motor**	**Cognitive**	**Receptive Language**	**Expressive Language**	**Fine Motor**	**Gross Motor**
ASQ-3			*n* = 186					*n* = 151		
Problem Solving	**0.106**	0.136	0.248***	0.069	0.018	**0.176***	0.236**	0.168*	0.118	0.104
Communication	0.254***	**0.224****	**0.465*****	0.244***	0.154*	0.166*	**0.253****	**0.403*****	0.043	0.149
Fine Motor	0.153*	0.109	0.230**	**0.203****	0.205**	0.123	0.098	0.150	**0.251****	0.127
Gross Motor	0.108	0.055	0.026	0.152*	**0.279*****	0.080	0.102	0.025	-0.029	**0.396*****
Personal-Social	0.073	0.034	0.148*	0.070	0.048	0.099	0.144	0.186*	0.198*	0.140
Denver-II			*n* = 184					*n* = 146		
Language	0.329***	**0.351*****	**0.510*****	0.262***	0.274***	0.335***	**0.434*****	**0.587*****	0.220**	0.281***
Fine Motor-Adaptive	**0.381*****	0.350***	0.292***	**0.325*****	0.356***	**0.301*****	0.303***	0.242**	**0.362*****	0.105
Gross Motor	0.338***	0.349***	0.234**	0.310***	**0.560*****	0.108	0.172*	0.160	0.264**	**0.483*****
Personal-Social	0.193**	0.132	0.110	0.243***	0.178*	0.181*	0.196*	0.266**	0.199*	0.203*
SFI & SFII (MacArthur)			*n* = 124					*n* = 97		
Receptive Language[a]	0.177	**0.230**	**0.149**	0.143	0.145	0.079	**0.194**	**0.342***	-0.094	0.158
Expressive Language[b]	0.241**	**0.261****	**0.441*****	0.138	0.251**	0.179	**0.390*****	**0.509*****	0.219*	0.176
BDI-2 (Battelle)			*n* = 148					*n* = 167		
Cognitive	**0.279*****	0.372***	0.385***	0.244**	0.226**	**0.448*****	0.401***	0.349***	0.364***	0.261***
Communication	0.281***	**0.335*****	**0.511*****	0.232**	0.211*	0.381***	**0.385*****	**0.480*****	0.272***	0.263***
Motor	0.195*	0.255**	0.305***	**0.353*****	**0.344*****	0.242**	0.236**	0.150	**0.323*****	**0.333*****
Personal-Social	-0.035	0.058	0.194*	-0.114	0.004	0.225**	0.227**	0.252***	0.078	0.107
Adaptive Skills	0.068	0.114	0.187*	0.175*	0.156	0.225**	0.255***	0.238**	0.283***	0.219**
WHO-Motor			*n* = 35					*n* = 36		
Gross Motor[c]	0.105	-0.037	-0.123	0.029	**0.675*****	0.087	0.095	0.353*	0.096	**0.443****

* p<0.05, ** p<0.01, *** p<0.001. Pearson correlations on internally standardised scores, Standard Errors (SE) computed using bootstrap stratifying by age category and socio-economic sector (*n* =1000 replications). Matching scales bolded. [a] Children 8-18 months; [b] Children 8-30 monts; [c] Children 6-15 months.

Appendix

Appendix I: Publisher Website[21], Language Adjustments and Other Adaptations Made to Test Items

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

The Bayley-III became available in Spanish in mid-2015. Hence, we translated and back-translated the English version of the test to Colombian Spanish.

In addition, we had to modify the following images in the language scales:
- In various items, for the action word 'washing' (*lavando*), we modified the image of a washing machine in the stimulus book by a regular washing sink. The action word remained the same.
- In various items, the action word 'vacuuming' (*aspirando*) was replaced by 'sweeping' (*barriendo*) and the image in the stimulus book was modified accordingly.

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

We modified the following words from those in the original version of the ASQ-3 in Spanish:
- Gross Motor Scale. Items 9 (questionnaire 33) and 8 (questionnaire 36): *Resbaladilla* was replaced by *rodadero*, a word more commonly used in Colombia.
- Fine Motor Scale. Items 3, 5, 7 (questionnaire 6), items 1, 3, 5, 8 (questionnaire 8), items 2, 5, 7 (questionnaires 9 and 10), items 2, 4 (questionnaire 12), item 1 (questionnaire 14): 'Cheerio' was replaced by *bolita de cereal*.
- Problem Solving Scale. Item 8 (questionnaire 8), item 5 (questionnaires 9 and 10), items 2, 7 (questionnaire 12), items 4, 8 (questionnaire 14), items 2, 6 (questionnaire 16), items 3, 6 (questionnaire 18), item 8 (questionnaire 20), item 5 (questionnaire 22) and item 2 (questionnaire 24): 'Cheerio' was replaced by *bolita de cereal*.
- Communication Scale. Items 9 (questionnaire 24) and item 6 (questionnaire 27). In order for it to become more logical, the instruction was modified from "*Pon el zapato encima de la mesa y pon el libro debajo de la silla*" to *"pon el libro sobre la mesa y pon el zapato debajo de la silla*".

Denver-II
Publisher: Denver Developmental Materials Inc.
http://denverii.com/

Some parts of the administration manual required translation to Spanish. In addition, we modified the following items in the record forms and test instructions:

[21] All publisher websites were last consulted on March 25, 2016.

- Items 1 and 7. Spelling mistakes were corrected.
- Items 9 and 11. The word *dedos* was added to clarify the instructions.
- Items 21 and 24. Instructions in the record form were modified so that they matched those in the test manual.
- Item 25. Replace *banana* with *banano* and *cerca* with *reja*, since these are more commonly used in Colombia.
- Item 26. Spelling mistakes were corrected, as well as instructions in the record form so as to match those in the test manual.

SFI

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

CDI Advisory Board
http://mb-cdi.stanford.edu/board.html

The original version was developed for Mexico. To ensure linguistic equivalence of the test items in Colombia, we made the following modifications to the test:

- *Guagua* was replaced by *guaguau*
- *Camión/troca* was replaced by *bus*
- *Coche* was removed as an option from *carro/coche*
- *Tortilla* was replaced by *arepa*
- *Botella/mamilla* was replaced by *tetero*
- The word *plata* was added as an option to dinero (*dinero/plata*)
- *Lavabo* was replaced by *lavamanos*
- The word *templo* was excluded as an option from *templo/iglesia*
- *Byebye* was replaced by *chao*

SFII

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

Similarly, and in addition to modify the words guagua, carro/*coche, camion/troca, botella/mamilla, iglesia/templo* and *byebye* (see modifications to SFI), we made the following modifications to ensure linguistic equivalence of the test items in Colombia:

- *Víbora was replaced by culebra/serpiente*
- *Plátano/banana was replaced by plátano/banano*
- *Calabaza was replaced by tomate*
- *Chícharo was replaced by pollo*
- *Cerillos was replaced by fósforos*

BDI-2

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

Even if we had the Spanish version of the test, the Item Test Book, which includes the specific instructions for accurate item administration and scoring, had to be translated from English to Spanish. Similarly, the text in the Picture Book had to be translated.

In addition, we had to modify translated items in the following manner to ensure functional comprehension amongst our study sample:

- Item 2: *mama* was changed to *alimentarse*.
- Item 3: *traga* was replaced for *pasa la comida*.
- Item 37: *centavos* (cents) were replaced by *pesos*.
- Item 80: *cordel* was modified to *pita*
- Items 10, 25, 26, 31, 332 and 94: instructions were modified since they were difficult to understand by parents and children as they were.
- Item 60: modified since they were based on illustrations that had words in English.
- Item 95: replaced picture of *train* by a picture of a *bus*.

WHO-Motor

Publisher: WHO

http://www.who.int/childgrowth/standards/motor_milestones/en/

The test and administration and scoring instructions were translated to Colombian Spanish.

Appendix II: Internal Standardization of Scores using Age-Conditional Means and SDs.

For each scale, we removed tester effects from the raw score by running a regression of the raw scores on tester dummies using Ordinary Least Squares (OLS). We constructed the residuals of these regressions, which we standardized by age using non-parametric methods as follows. First, we computed the age-conditional mean using the fitted values of the regression in (1), estimated by kernel-weighted local polynomial smoothing methods:

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

where $\mathrm{Y_i}$ is the residual of the raw score of child *i* in a given scale of a regression on tester dummies. X_i is the age of the child in days. Next, we regressed the square of the residuals in (1) on age of the child (in days) as shown in the kernel-weighted local polynomial regression in (2):

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

Our estimate of the age-conditional standard deviation (SD) is the square root of the fitted values $\hat{\mathrm{g}}_i$ in (2). Finally, we computed the internally age-adjusted z-score, $\mathrm{ZY_i}$, by subtracting from the residual of the raw score the within sample age-conditional mean estimated in (1) and dividing by the within sample age-conditional SD obtained from (2). More specifically:

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

This resulted in smooth normally distributed internally standardized scores, with mean zero across the age range (available upon request).

Appendix III: Appendix Tables

Table A1: Bayley-III Raw and Composite Scores and Short Tests Raw Scores - Overall and by Age

	6-42 mths		6-18 mths		19-30 mths		31-42 mths	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Bayley-III Raw Scores	(*n* = 1311)		(*n* = 443)		(*n* = 454)		(*n* = 414)	
Cognition	58.74	14.06	42.38	8.04	61.59	5.81	73.13	4.07
Receptive Language	25.44	9.12	15.27	3.46	26.36	5.14	35.32	3.49
Expressive Language	25.25	10.27	14.30	4.03	25.64	5.04	36.53	5.71
Fine Motor	39.32	10.07	28.44	5.30	39.63	3.95	50.61	4.45
Gross Motor	52.38	11.57	39.10	9.02	55.62	3.66	63.05	2.76
Bayley-III Composite Scores								
Cognition	98.37	8.91	103.86	9.44	95.73	7.93	95.40	6.25
Language	96.49	9.88	99.27	9.95	93.30	10.14	97.02	8.45
Motor	99.55	10.85	95.79	11.55	99.47	10.08	103.66	9.33
ASQ-3 (9 items)	(*n* = 664)		(*n* = 221)		(*n* = 224)		(*n* = 219)	
Problem Solving	46.86	15.27	46.77	12.93	46.59	15.93	47.24	16.75
Communication	46.73	18.13	44.41	14.34	42.99	17.76	52.90	20.27
Fine Motor	46.61	15.19	45.68	15.38	44.49	13.14	49.73	16.47
Gross Motor	50.14	17.52	44.93	20.56	51.81	14.20	53.68	16.04
Personal-Social	48.06	14.46	47.87	14.91	44.98	11.93	51.39	15.67
Denver-II	(*n* = 658)		(*n* = 225)		(*n* = 221)		(*n* = 212)	
Language	20.41	6.32	13.86	2.82	20.83	3.29	26.93	4.01
Fine Motor-Adaptive	18.77	4.39	13.81	2.60	19.73	1.91	23.03	1.89
Gross Motor	20.90	5.48	14.43	3.49	22.69	1.99	25.89	1.83
Personal-Social	15.93	5.04	10.09	3.23	17.35	1.75	20.65	1.97
SFI (MacArthur)[a]	(*n* = 192)		(*n* = 192)					
Receptive Language	44.56	20.85	44.56	20.85				
Expressive Language	6.39	6.98	6.39	6.98				
SFII (MacArthur)[b]	(*n* = 226)				(*n* = 226)			
Expressive Language	52.44	26.18			52.44	26.18		
BDI-2 (Battelle)	(*n* = 635)		(*n* = 215)		(*n* = 227)		(*n* = 193)	
Cognitive	16.93	4.01	13.27	2.66	17.03	1.70	20.89	3.30
Communication	17.40	6.57	10.54	4.09	18.16	2.63	24.15	4.09
Motor	18.52	6.38	11.21	4.49	20.36	2.52	24.51	2.15
Personal-Social	18.76	5.79	13.32	3.85	19.54	3.41	23.89	4.47
Adaptive Skills	16.39	5.51	10.46	3.22	17.77	3.03	21.39	3.32
WHO-Motor[c]	(*n* = 152)		(*n* = 152)					
Gross Motor	3.99	2.08	3.99	2.08				

[a] Children 8-18 months; [b] Children 19-30 months; [c] Children 6-15 months.

Table A2: Correlations between the Short Tests and the Bayley-III Composite Scores, by Age Group

	Bayley-III, 6-18 months			Bayley-III, 19-30 months			Bayley-III, 31-42 months		
	Cognitive	Language	Motor	Cognitive	Language	Motor	Cognitive	Language	Motor
ASQ-3		*n* = 221			*n* = 224			*n* = 219	
Problem Solving	**0.161***	0.089	0.024	**0.063**	0.130	0.081	**0.307*****	0.441***	0.236***
Communication	0.143*	**0.230*****	0.191**	0.168*	**0.404*****	0.076	0.384***	**0.547*****	0.321***
Fine Motor	0.024	0.035	**0.067**	0.109	0.155*	**0.184****	0.308***	0.299***	**0.401*****
Gross Motor	0.095	0.038	**0.368*****	-0.072	0.013	**0.109**	0.186**	0.084	**0.215****
Personal-Social	0.195**	0.188**	0.153*	0.000	0.093	0.089	0.060	0.198**	0.139*
Denver-II		*n* = 225			*n* = 221			*n* = 212	
Language	0.051	**0.191****	0.054	0.228***	**0.536*****	0.148*	0.490***	**0.616*****	0.406***
Fine Motor-Adaptive	**0.166***	0.083	**0.139***	**0.344*****	0.367***	**0.314*****	**0.387*****	0.387***	**0.456*****
Gross Motor	0.092	0.016	**0.372*****	0.005	0.150*	**0.274*****	0.196**	0.210**	**0.323*****
Personal-Social	0.196**	0.159*	0.137*	0.054	0.153*	0.210**	0.231***	0.158*	0.112
SFI & SFII (MacArthur)		*n* = 192[a]			*n* = 226				
Receptive Language	0.140	**0.181***	0.103						
Expressive Language	0.035	**0.178***	0.029	0.179**	**0.472*****	0.116			
BDI-2 (Battelle)		*n* = 215			*n* = 227			*n* = 193	
Cognitive	**0.253*****	0.173*	0.273***	**0.205****	0.307***	0.180**	**0.490*****	0.459***	0.324***
Communication	0.133	**0.065**	0.228***	0.302***	**0.513*****	0.192**	0.422***	**0.630*****	0.332***
Motor	0.077	0.018	**0.298*****	0.218***	0.237***	**0.331*****	0.355***	0.309***	**0.388*****
Personal-Social	0.058	0.120	0.101	0.157*	0.295***	-0.046	0.127	0.226**	0.051
Adaptive Skills	0.031	0.189**	0.180**	0.066	0.186**	0.198**	0.220**	0.286***	0.171*
WHO-Motor		*n* = 152[b]							
Gross Motor	0.190*	0.149	**0.513*****						

* p<0.05, ** p<0.01, *** p<0.001. Pearson correlations between Bayley-III composite scores and internally standardised scores of the short tests. Standard Errors (SE) computed using bootstrap stratifying by age category and socio-economic sector (*n* =1000 replications). Matching scales bolded.

[a] Children 8-18 months; [b] Children 6-15 months.