Tracking early child development at the population level: Validation of the Kidsights Measurement Tool for children birth to age five years

Marcus R. Waldman (marcus.waldman@unmc.edu)
University of Nebraska Medical Center
Katelyn Hepworth
University of Nebraska Medical Center
Jolene Johnson
University of Nebraska Medical Center
Kelsey M. Tourek
University of Nebraska Medical Center
Kelly J. Jones
University of Nebraska Medical Center
Yaritza Estrada Garcia
University of Nebraska Medical Center
Laura M. Fritz
University of Nebraska Medical Center
Abbey Siebler
Abbie Raikes
University of Nebraska Medical Center

Research Article

Keywords:

Posted Date: July 5th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3084382/v1

Abstract

Background

Group disparities in early child development arise due to inequities in early environments that are reflective of socio-economic status, geography, and other factors. To track and address these disparities, valid and reliable child development tools are needed that can be implemented at-scale and across populations. However, no population-based measures of child's motor, cognitive, language, and social/emotional development appropriate for children from birth to age five years have been validated in the United States to date. In response, we have designed the Kidsights Measurement Tool (KMT).

Methods

We evaluate the validity and reliability evidence of the KMT with reference to the *Standards for Educational and Psychological Testing* (APA, AERA, & NCME, 2014) from a sample of $N = 3{,}413$ initial parent reports residing in Nebraska, USA, as well as from a 12–24-month direct observation follow-up subsample of $N = 70$ children. Evidence came from the test content, evidence based on other variables, and the precision of scores.

Results

We find strong evidence supporting the KTM's validity and reliability ($r_{XX'} = .92$) as a population-based tool. We find that scores from KMT's initial administration strongly predict later scores from the Bayley Scales of Infant Development ($r > .50$) and the Woodcock Johnson's overall development score ($r = .70$), both administered by trained professionals at least one year later. We also find that scores exhibit expected associations with known correlates of children's development, including the parent's educational attainment, enrollment into governmental subsidies and services, parent's anxiety and depressive symptoms, and the child's count of adverse childhood experiences.

Background

Early child development lays the groundwork for lifelong health and wellbeing. Supportive early environments that promote healthy child development are associated with later adult health and well-being, including lower incidence of diseases, greater earnings, and lower risk of incarceration (Campbell et al., 2014; Cohen et al., 2010; Reynolds et al., 2007; Walker et al., 2022). A substantial body of research outlines the biological (Feder et al., 2019; Gluckman et al., 2006) and social mechanisms (Garner et al., 2021; Shonkoff et al., 2021) by which early development influences later health and wellbeing. These mechanisms include exposure to excessive stress that impedes development of self-regulatory skills, contributes to chronic inflammation that in turn leads to lifelong poor health, and interferes with cognitive development (Shonkoff et al., 2012). Moreover, the impacts of early stress are visible in group-level disparities in health, learning, and wellbeing throughout the life course (Marmot et al., 2005). Because early child development has such significant implications for later development, investments in early childhood development yield notable long-term returns (Heckman et al., 2006), in part because supporting healthy development early in life is more cost-effective than mitigating the consequences of early stress and deprivation in adulthood.

Children's development is strongly influenced by the social and economic context of their families and communities and, when these contexts are not supportive, lead to persistent group-level disparities in child development outcomes. Population disparities in cognitive, language, and social/emotional development outcomes in young children by socio-economic status, race/ethnicity, and geography have been extensively documented over several decades in the United States (e.g., MacLeod, 1998; Shonkoff & Phillips, 2000), emerging in the first year of life and persisting over time (Halle et al., 2009; Burchinal et al., 2011; Duncan et al., 2005). These disparities have been shown to be attributable to the differences in environmental supports, including access to quality childcare, economic resources, and neighborhood and community supports (Burchinal, et al., 2011; Duncan et al., 2005).

Population-level tracking on child development: Too little, too late

Population tracking of early development, defined by collecting and using data that can produce valid estimates of population disparities (i.e., data collected from representative samples of children birth to age five years from valid and reliable early childhood tools) is limited in scope in the United States. To date, population tracking of group-level disparities for young children has relied on a small and narrow set of indicators, especially for children birth to age three. Infant mortality, for example, is a common indicator of early inequity (e.g., Dodge, 2022), but infant mortality does not provide insight into young children's development after birth. Using early childhood tools administered after birth, a few studies have found that children from lower-income families, families with less formal education, and/or families who represent diverse racial and ethnic backgrounds are up to half a year behind their more advantaged peers in cognitive, language, and social/emotional development by the time formal schooling begins (Friedman-Krauss et al., 2020; Ghandour et al., 2021). Given the substantive size of these disparities, more population-level data are needed on children's developmental outcomes earlier in life.

Moreover, US state and federal agendas for population-level health and wellbeing include key indicators of early childhood development. Healthy People 2030, for example, states an objective of increasing the national proportion of children who are developmentally on track and ready to start school (CDC, 2023). But progress toward this objective is not currently reported, as the CDC has designated the Healthy People 2030 goal regarding children's developmental status as "presently lacking reliable baseline data" (CDC, 2023). The landscape of routinely collected population-based child development data for children birth to age five years in the United States includes one measure. The National Survey of Children's Health, administered by the Health Resources and Services Administration using a representative sampling design, includes a National Outcomes Measure on preschool-aged children's development to generate population-level estimates, called Healthy and Ready to Learn (HRTL; Ghandour et al., 2019). The percentage of children who are HRTL is estimated based on a set of 22 parent-report items to index child development across physical/motor, early learning skills, self-regulation, and social/emotional development for children between three and six years (Ghandour et al., 2019). Results from national samples estimate that only 42% of children are HRTL in all developmental domains. Significant disparities were found by—among other factors—the parent's mental health status, the quality of the home learning environment, and the

child's neighborhood (Ghandour et al., 2021). However, HRTL does not collect data of children's development for children under age three years, well after disparities emerge. To summarize, despite the policy goals of collecting population data on infant and toddler wellbeing, data on young children's social/emotional, language, and cognitive development beginning at birth are largely absent from statewide data systems in the United States (Ryberg et al., 2022).

Clearly, more data are needed to provide important insights into population-level disparities in child development in the first years of life. The lack of data is especially problematic given the rapid pace of early development and the opportunity to support healthy development through cost-effective programs such as home visiting and support for quality childcare (Shonkoff & Phillips, 2000; Jeong et al., 2021).

Status of population-based measurement

As noted previously, assessment of progress towards national and state goals requires data generated with population-based samples, or those that are representative of the underlying population from valid and reliable early childhood tools. Reliable data on multiple aspects of child development is a critical piece of informing effective policies and programs, and population-level data on children becomes increasingly important as the scale and scope of programs expand (Yoshikawa et al., 2018; Ryberg et al., 2022). Results from population-based tools of child development are primarily intended to inform policy by identifying groups of children in need of additional support, tracking progress towards national and state goals, and providing an estimate of the impacts of large-scale, community or state-based early childhood programs (Ryberg et al., 2022), rather than identifying specific children in need of additional services. Of course, indicators of child development can and should be viewed in the context of other information on children's health, economic wellbeing, access to childcare and other factors, which describe the contextual and environmental determinants of young children's development (Paschall et al., 2020), and when taken together, create an overall picture of early disparities within the population of young children in a given geographic area. Moreover, ideally, population-based tools of child development capture multiple domains of early development, including language, cognitive, motor, and social/emotional development, as all domains contribute to lifelong health and wellbeing. In summary, population-level indicators of child development should generalize to the underlying population, support valid inferences regarding young children's development as early as disparities emerge, including demonstrating sensitivity to factors associated with disparities in child development.

Measurement of population-level trends in child development requires measurement tools that are feasible to use at scale in that they are cost-manageable and do not require undue resources and time to implement. Thus, the considerable cost of direct assessments of child development, such as the Bayley Scales of Infant Development, preclude their use at a population level. Instead of reliance on direct assessments of children, parent-reported measures administered using surveys can be used at scale to estimate group disparities in child development, particularly when administered through online surveys that facilitate fast and easy data collection.

While only population-based measure of early child development is routinely used in the United States, HRTL, three measures have been designed to measure child development for children under age five at the global population level: the Caregiver Reported Early Development Inventory for children up to age three years (CREDI; McCoy et al., 2018), the Early Care and Development Index designed for children two to five years as part of UNICEF's global monitoring agenda (ECDI2030; Halpin et al., 2023), and the Global Scale for Early Development for children birth to age three years (GSED; Cavallera et al., 2023). As results are intended to inform policy, it is important to address the validity of population-based instruments of child development. The *Standards for Educational and Psychological Testing* (henceforth, *Standards*; American Psychological Association [APA], American Educational Research Association [AERA], & National Council of Measurement in Education [NCME], 2014) define multiple types of validity evidence: (1) evidence from test content, (2) evidence from relations with other variables (including criterion variables and scores from concurrent instrument), (3) the sensitivity of conclusions to validity threats (e.g., measurement invariance or item misfit), and (4) the precision of scores (i.e., reliability). To our knowledge, few studies to date have comprehensively examined the psychometric properties of child development measures intended for use at the population level, despite the importance of generating population-level estimates of group disparities among young children. McCoy et al. (2018) reported acceptable psychometric properties and criterion validity of the CREDI across high, middle and low-income countries. Ghandour et al. (2019) similarly reported acceptable psychometric properties and criterion validity for HRTL. Validation evidence for the GSED and ECDI are still in development. However, two types of validity evidence – concurrent validity with observational measures and predictive validity demonstrating associations between parent-report and observational measures over time – have not yet been frequently reported for population-based measures of child development for children birth to age five years.

The purpose of the present study was to evaluate the validity evidence of a new parent-report measure of child development for children birth to age five living in the United States called the Kidsights Measurement Tool (KMT). The KMT was developed to generate population-based estimates of children's development from birth through age five, to address the present gap in holistic indicators of child development with a measure that is feasible to scale. We hypothesized that the KMT would show acceptable psychometric properties, based on the existing literature on population-based instruments to measure child development. Using the *Standards*, we assessed the KMT's validity using a large sample from Nebraska, USA, including predictive validity with direct assessment administered by trained observers 12–24 month later. We also assessed criterion validity with known correlates of children's development, including parent education, parent anxiety and depression, the family's socio-economic status, and exposure to experiences that may be traumatic and/or adverse to healthy development.

Methods

Participants

In total, the present study included initial ("Time 1") responses from the parents of N = 3,413 eligible children 0–71 months residing in Nebraska, USA. Additionally, both parent report and observational data also came from a 12–24 month follow-up subsample ("Time 2") of N = 70 children who were 0–47

months old at Time 1. At Time 1, 53.3% of children were male, 55.9% were identified as white, non-Hispanic, the mean age was 34.6 months. At Time 2, 42.9% were male, 58.6% were identified as white, non-Hispanic, and the mean age was 37.8 months.

The complete set of inclusion and exclusion criteria for participant eligibility is diagramed in the study identification flowchart shown in Fig. 1. To be eligible for the present study, participants must be parents of children aged birth to five years, residing in Nebraska. “Parents” were defined as adults who were responsible for the child’s care at least 40 hours a week and identified themselves as biological, foster, or adoptive parents, or other relatives. At Time 1, 97.6% of respondents reported that they were the biological, foster, or adoptive parent of the child; at Time 2 all respondents (100%) identified as the biological, foster, or adoptive parent. Throughout this paper, we refer to respondents as “parents.” Because the present study focuses on the validation of the English version of the KMT, we excluded all responses to the Spanish version. The Time 2 sample was identified from the respondents to the Time 1 survey based on their willingness to be contacted again. All Time 2 sample participants completed the Time 1 survey before May 2021.

Studies have noted that financial incentives (like those we offered our participants) increase the likelihood of receiving fraudulent responses (c.f., Lawor et al, 2021). As part of a screening protocol, we excluded observations if (a) metadata information resulted in a “likely fraudulent” score from the *rIP* package (Waggoner et al, 2019) and the IP Hub database (https://iphub.info/), (b) the caregiver failed to accurately confirm the child’s birthdate, and (c) whether scores were above or below 5SD on the CREDI or ECDI (see below for a description of each). We refer to all initial administrations that met the eligibility criteria as the Time 1 sample (N = 3,413).

The duration between the initial administration at Time 1 and the follow-up at Time 2 averaged M = 16 months (range = 12–24). We excluded from analysis any observations which did not meet basal requirements for the direct assessments administered at Time 2. We refer to eligible response at follow-up as the Time 2 sample (N = 70).

Procedures

Data for Time 1 responses were collected using an online survey in between October 2020 and February 2023. Participants were given the option to take the survey in English or in Spanish. We offered parents a gift card ($20 to $40) to complete the survey. We recruited parents through healthcare providers, childcare and parenting support programs, and social media posts. We gave parents a link to an online questionnaire including several questions on family demographics and the child’s adverse childhood experiences (ACEs; described below), development, health, and home environment. Respondents could complete the survey using their mobile phone, tablet, or computer and took between 20 to 30 minutes to complete.

Measures

Kidsights Measurement Tool

The Kidsights Measurement Tool was constructed by first forming a candidate item bank by adopting items from four previously validated instruments each measuring normative aspects of children’s development (i.e., skills or behaviors that children acquire or exhibit as they age when undergoing healthy development) between 0–5 years. These four instruments included (1) the Global Scale of Early Development Short Form (GSED-SF; McCray et al., 2023), (2) the Caregiver Reported Early Development Instruments Long Form (CREDI-LF; McCoy et al. 2018), (3) the Early Childhood Development Index (ECDI2030; Cappa et al., 2021), and (4) Healthy and Ready to Learn (HRTL; Ghandour, 2019). We included only items that measured normative aspects of children’s development (i.e., skills or positive behaviors that are acquired or manifest as children age under healthy development), and we excluded items from these instruments that measure constructs such as problem behaviors or other indicators of psychosocial difficulties.

This process resulted in a candidate item bank of 223 items with 79 items unique to the GSED-SF, 23 items unique to the CREDI-LF, 7 items unique to the ECDI2030, and 49 items unique to the HRTL. Of the 223 items, 49 items were shared across one or more of the four contributing instruments (42 items were common between GSED-SF and CREDI; 7 items were common between the GSED-SF, CREDI-LF, and ECDI2030).

The 223 candidate items measure motor, cognition, language, and/or social/emotional constructs according to the published literature and existing documentation for the four instruments. Specifically, 71 items represented fine or gross motor development constructs (c.f., McCray, 2023; *redacted*, 2021; and Cappa et al., 2021) or physical development (c.f., Ghandour, 2019). Additionally, 82 represented cognitive or language development (c.f., McCray, 2023; *redacted*, 2021; and Cappa et al., 2021) or early learning skills (c.f., Ghandour, 2019). Lastly, 70 items measure social/emotional development (McCray, 2023; *redacted*, 2021; Cappa et al., 2021; Ghandour, 2019) including normative aspects of children’s self-regulation (Ghandour, 2019).

The 223 candidate items were then screened for sufficient variability in responses, including at least a 90% endorsement probability at birth. This process led to 19 items being removed from the candidate pool (see Supplemental Table 1). The result was a final set of 204 items spanning development from birth to age 5 years.

Concurrent and Predictive Measures

We administered previously validated direction-observation measures at Time 2 and parent-reported instruments at both Time 1 and Time 2.

Direct-observation instruments. Two direct assessments were administered: (a) The Bayley Scales of Infant and Toddler Development, Fourth Edition (Bayley-4; Bayley & Aylward, 2019) and (2) the Woodcock Johnson IV Early Cognitive and Academic Development (WJ IV ECAD; Schrank et al., 2018). The Bayley-4 and the WJ IV ECAD were only administered to follow-up subsample at Time 2 (N = 70). The WJ IV ECAD was used for children from 43–60 months of age at Time 2 (n = 33), and the Bayley-4 was used for children up to 42 months at Time 2 (n = 37).

Bayley-4. The Bayley-4 is validated to measure child development up to 42 months (e.g., Klein-Radukic, et al., 2023). The instrument is divided into items that capture development in the cognitive, language, motor, social/emotional and adaptive behavior domains through direct administration of activities,

observation of the child, and questions to the caregiver (Bayley & Aylward, 2019). Scores are provided at domain level and the subtest level.

For the Bayley-4 training, assessors were required to complete a 12-hour online training hosted by the measure publisher. After completing the training, assessors submitted video recordings of administrations of the Bayley-4 or scheduled in-person observations with the research team's Bayley-4 supervisor. The measure supervisor has several years of experience administering the Bayley Scales of Infant Development and evaluated each assessor for correct administration of item and scoring in order to certify each assessor as reliable on the Bayley-4.

WJ IV ECAD. The WJ IV ECAD is considered a measure of intellectual ability, academic skills and language, specifically oral expression for children 30 months to 6 years old (Schrank et al., 2018). The results of the WJ IV ECAD administrations result in a General Intellectual Ability Score, an Expressive Language Score, and scores by each test (LaForte et al., 2015). Although there are 10 tests in the WJ IV ECAD, only 7 of the tests were administered for the study. For this study, only the Bayley-4 scales cognitive, language and motor scales were administered to children up to 42 months.

Assessors followed a similar training and reliability process for the WJ IV ECAD as for the Bayley-4. They reviewed the WJ IV ECAD kit materials and submitted video of the administration of the WJ IV ECAD. The research team's WJ-ECAD supervisor has experience administering the measure and reviewed the recording for correct administration and scoring before certifying the assessors as reliable on the measure.

Caregiver-reported instruments. The candidate Kidsights item pool included all items from the GSED-SF, CREDI-LF, ECDI2030, and HRTL. As a result, in administering the KMT, we effectively administered these four caregiver instruments concurrently. Scores from the GSED-SF are termed "D-scores" (Weber et al., 2019) and calculated using the *dscore* (van Buuren, Eekhout, & Huizing, 2022) R package. The CREDI Long Form results in an overall score of child development as well as subscale scores of motor, cognition, language and social/emotional development (Seiden et al, 2021; *citation redacted*). We calculated CREDI scores using the *credi* R package (https://github.com/marcus-waldman/credi). ECDI2030 scores were calculated using UNICEF's (2023) provided R syntax file. HRTL scores were calculated by replicating the four-factor solution reported in Ghandour (2019) using the *lavaan* (Rosseel, 2012) R package and extracting factor scores. The four factors include a physical/motor factor, an early learning factor, a social/emotional factor, and a self-regulation factor.

Global Scales of Early Development Psychosocial Form. In addition to the KMT, we administered the Global Scales of Early Development (GSED-PF; *citation redacted*). The GSED-PF is currently undergoing validation and measures manifestations of early psychosocial stresss. The GSED-PF includes an overall score of children's behaviors, as reflected through internalizing behaviors, externalizing behaviors, feeding problems, sleeping difficulties, and social competency problems.

Family- and Caregiver-Level Criterion Measures

Socioeconomic Measures. Using questions taken from the National Survey of Children's Health, we asked parents to report their enrollment in governmental services and programs, educational attainment, and household income. Parents were asked to report if they were enrolled in 1) Medicaid; 2) Cash assistance from a government welfare program; 3) Free or reduced-cost breakfasts or lunches at school; 4) Food Stamps or Supplemental Nutrition Assistance Program (SNAP) benefits; 5) Benefits from the Woman, Infants, and Children (WIC) Program. For analysis, we created a single dummy variable (denoted GOVT) indicating whether the caregiver was enrolled in any of the above governmental programs (1-Yes, 0-No).

In the 2020 survey, we asked parents to report their 2019 household income in United States Dollars (USD). Likewise, in the 2022 survey, we asked parents to report their 2021 household income. To make household income on the same scale across years, we adjusted for inflation by converting to 1999 USD.

Parents could select from nine options in reporting their educational attainment: 1) 8th grade or less; 2) 9th-12th grade; 3) No diploma; 4) High School Graduate or GED Completed 5) Completed a vocational, trade, or business school program 6) Some College Credit, but No Degree 7) Associate Degree (AA, AS); 8) Bachelor's Degree (BA, BS, AB); 9) Master's Degree (MA, MS, MSW, MBA); 10) Doctorate (PhD, EdD) or Professional Degree (MD, DDS, DVM, JD). In the present study, we collapsed this information into four categories including 1) no high school (HS) diploma; 2) HS diploma; 3) Some college or an Associate's degree (i.e., AA/AS); 4) Bachelor's degree (i.e., BA/BS) or higher. We dummy coded caregiver educational attainment using parents with only high school education as the reference group.

Anxiety and Depressive Symptoms. We administered the Patient Health Questionnaire 2-item (PHQ-2; Kroenke et al., 2003) and the Generalized Anxiety Disorder 2-item (GAD-2; Löwe et al., 2008) to obtain caregiver self-reports of depressive and anxiety symptoms. Parents reported whether, over the last two weeks, they (1) had little pleasure or interest in doing things (i.e., indicator 1 of PHQ-2) and (2) were feeling down, depressed, and hopeless (i.e., indicator 2 of PHQ-2), (3) were feeling nervous, anxious, or on edge (i.e., indicator 1 of GAD-2), and (4) were not able to stop or control worrying (indicator 2 of GAD-2). Parents responded using a four-point Likert scale (i.e., "0-Not at all"; "1-Several days"; "2-More than half the days"; "3-Nearly every day"). For analysis, we created a depression and anxiety symptom total score by summing all four items.

Child-Level Criterion Measures.

Child-level criterion variables included information on the child's sex (i.e., male or female), race, ethnicity, overall health status (as reported by the caregiver), and exposure to adverse childhood experiences (ACEs; Felitti et al., 1998). We adopted the survey items from the NSCH (Ghandour et al., 2018) in collecting this information.

Race and Ethnicity. Parents could select from up to 15 racial categories and one of five ethnicity categories. Racial category response options included: 1) American Indian or Alaska Native; 2) Asian Indian; 3) Black or African American; 4) Chinese; 5) Filipino; 6) Guamanian or Chamorro; 7) Japanese; 8) Korean; 9) Native Hawaiian; 10) other Asian; 11) other Pacific Islander; 12) Samoan; 13) Vietnamese; 14) White; or 15) Some other race. Ethnicity response options

included 1) No, not of Hispanic, Latino, or Spanish origin; 2) Yes, Mexican, Mexican American, Chicano; 3) Yes, Puerto Rican; 4) Yes, Cuban; 5) Yes, another Hispanic, Latino, or Spanish origin.

For analysis, we combined racial and ethnicity into four major categories. These included: 1) White, non-Hispanic, 2) Black or African American, non-Hispanic, 3) Other (including two or more races), non-Hispanic, and 4) Hispanic.

Child's General Health. We asked parents to rate their child's general health ("In general, how would you describe this child's health?"). Response options included: 1) Poor, 2) Fair, 3) Good, 4) Very Good, or 5) Excellent. For analysis, we applied dummy coding to indicate whether the child was reported to be in very good or excellent health (1-Yes, 0-No).

Adverse Childhood Experiences. We administered eight items measuring children's ACEs all of which include "0-No" or "1-Yes" as response options: 1) caregiver divorce or separation, 2) caregiver death, 3) a household member with a drug or alcohol problem, 4) caregiver diagnosed with a mental illness, 5) exposure to violence in the community, 6) exposure to domestic violence, 7) parental incarceration, 8) racism. We determined the child's count ACEs by summing the responses to the eight items.

Scaling and Scoring Procedures

Se fit the graded-response IRT model in (1)-(2) to the polytomous data.

$$\Pr(Y_j \geq k) = \text{logit}^{-1}\left(\alpha_j(\theta_i - \delta_{jk})\right), \qquad 0 < k \leq K_j \tag{1}$$

$$\theta_i = \gamma_0 + \gamma_1 \text{AGE}_i + \gamma_2 \text{AGE}_i^2 + \gamma_3 \text{AGE}_i^3 + \gamma_4 \text{AGE}_i^4 + \eta_i, \qquad \eta_i \sim N(0,1) \tag{2}$$

where i indexes a child (with a latent score [i.e., ability] of θ_i), j indexes an item (with K_j response options), and k indexes one of the responses options for the item. Model parameters in (1)-(2) include α_j (the item discrimination value), δ_{jk} (the difficulty value associated with response option k for item j), and the vector of latent regression coefficients γ. We fit the model using maximum marginal likelihood estimation (also referred to as full information maximum likelihood). Maximum likelihood estimators are gold standard approaches to treating missing data (Schafer & Graham, 2002). All model fitting occurred using the MIRT (Chalmers, 2012) package in R (R Core Team, 2022). We calculated Kidsights scores by summarizing the posterior distribution using the expected-a-posteriori (EAP) point estimate.

Analytic Plan

We followed the *Standards* in collecting and analyzing validity evidence. Evidence came from analyzing (1) test content, (2) relations with other variables (including criterion variables and scores from concurrent instrument), (3) the sensitivity of conclusions to threats from measurement non-invariance or item misfit, and (4) score precision as indicated by errors of measurement (i.e., reliability).

Evidence Based on Test Content

Using the domain assignments provided by the originating instruments, we designated items into one of three domains: (1) Motor/physical development, (2) Cognition or language development, or (3) Social/emotional development. To assess content coverage, we calculated the (average) domain composition of the administered items within yearly age categories. Evidence based on test content was evaluated by two subject matter experts to ensure adequate representation of items by developmental domain.

Evidence Based on Other Variables

In line with the *Standards*, we collected evidence that Kidsights scores correlate with other variables in the expected magnitude and direction. This includes: (a) convergent validity evidence with scores from concurrent instruments that measure equivalent (or highly similar) constructs as those directly intended to be measured by the KMT; (b) discriminant validity evidence with scores from concurrent measures measuring constructs *not* directly intended to be measured using the KMT; and (c) association of Kidsights scores with exogenous criterion variables known to be predictive of child development.

Convergent Validity Evidence. Convergent validity evidence from calculating part correlations (i.e., correlations after adjusting for the child's age) of Kidsights score with (a) the Bayley-4 Cognition, Receptive and Expressive Communication, and Gross and Fine Motor domain and subtest level scores for children 42 months and younger; and (b) the WJ IV ECAD General Intellectual Ability- Early Development scores, Expressive Language cluster scores, as well as the Verbal Analogies, Sentence Repetition, and Rapid Picture Naming test activities for children 43–60 months.

Convergent validity evidence from caregiver-reported instruments came from part correlations with: (a) D-scores, (b) CREDI scores (overall scores, as well as motor, language, cognition, and social/emotional subscores), (c) ECDI2030 scores, and (d) motor/physical, early learning, and social/emotional factor scores from the HRTL.

We are aware of no published ideal or minimum threshold for a correlation to establish convergent validity evidence. Because a correlation value of approximately $r = .70$ represents 50% of the variance explained, we used this value as an ideal threshold and a correlation of $r = .50$ as a minimum threshold (i.e., 25% of the variance explained).

Discriminant Validity Evidence. Discriminant validity evidence from part correlations (i.e., correlations after controlling for children's age) with scores from other instruments that reflect aspects of children's behaviors which a not directly tied to normative aspects of children motor, cognitive, language, or social/emotional development. This included scores from the HRTL self-regulation factor and psychosocial problem scores from the GSED-PF. Because these

scores reflect constructs not intended to be directly measured by the KMT, evidence comes from correlations smaller in magnitude (i.e., $|r|$<.50) and in the expected direction (i.e. negative part correlations with psychosocial problem scores, and positive correlations with self-regulation).

Predictive Validity Evidence. We assessed predictive validity evidence by studying part correlations of Kidsights scores at Time 1 with Bayley-4 and WJ-ECAD scores obtained 12–24 months (*M* = 16 months) later at Time 2. We took positive and statistically significant correlations as evidence that Kidsights scores predict future development and learning.

Criterion Associations. Following a model building procedure, we fit six multiple regression models (Models 1–6) of increasing complexity to evaluate criterion associations with variables known to be predictive of early childhood development. We controlled for child's age using a fourth-order polynomial so that the regression equations took the general form

$$\theta_i = \alpha_0 + \alpha_1 \text{AGE}_i + \alpha_2 \text{AGE}_i^2 + \alpha_3 \text{AGE}_i^3 + \alpha_4 \text{AGE}_i^4 + x_i\beta + \epsilon_i, \epsilon_i \ N(0, \sigma^2) \# \ (1)$$

where θ_i is the Kidsights score for the *i*th child and the criterion variables are elements in x_i.

In Model 1, we included a dummy variable indicating the child's sex as female (FEMALE; 0-Male; 1-Female) and tested whether females demonstrate higher average scores. In Model 2, we augmented Model 1 with an indicator variable for whether or not the caregiver reported the child was in very good or excellent overall health (HEALTHY; 0 – "Poor", "Fair", or "Good"; 1 – "Very Good" or "Excellent"). Building on Model 2, we specified Model 3 with three indicator variables of the caregiver's education attainment; we chose attaining a high school diploma as the reference category resulting in three dummy variables indicating that the caregiver had (a) not attained a high school diploma (NOHS; 0-No, 1-Yes), (b) attended college or attained an Associate's degree but had not earned a Bachelor's degree (SOMECOLL; 0-No; 1-Yes), or (c) attained Bachelor's degree or higher (denoted BS; 0-No, 1-Yes). We expected that increased educational attainment to be associated with higher average scores. In Model 4, we evaluated whether the caregiver's reported enrollment in governmental services (GOVT; 0 – Not enrolled; 1 – Enrolled in SNAP, WIC, FRPL, a cash assistance program, or a governmental healthcare program) were negatively associated with scores. In Model 5, we included information on the caregiver's race and ethnicity by specifying indicator variable for whether the caregiver was white, non-Hispanic (WHITE; 0-No, 1-Yes), Black, non-Hispanic (BLACK; 0-No, 1-Yes), or Hispanic (HISP; 0-No, 1-Yes); non-Hispanic parents identifying as two-or more races or who identified in a racial category other than black or white served as the reference category. After adjusting for differences in child's overall health, the caregiver's educational attainment, and enrollment in governmental services, we do not expect to find that race and ethnicity predict scores. Lastly, we included factor scores measuring the caregiver's depression and anxiety symptoms (DEPANX) and the child's adverse experiences (ACEs) in Model 6.

Missingness was most present in the GAD/PHQ-2 items (Min = 14.9%; Max = 15.3%), the survey question inquiring on the child's general health (14.5%), and the child's ACEs (with 0.50% not responding to at least one of the ACE questions). In response, we employed multiple imputation using the *mice* (van Buuren et al., 2015) R package to deal with missingness on the criterion variables. For all models, we evaluated the evidence from criterion associations by pooling results using Rubin's rules, conducting pairwise t-tests, and interpreting the magnitude and substantive size of the coefficients. For models with multiple coefficients to be tested (i.e., Models 3, 5, & 6), we conducted a simultaneous testing of the coefficients using a multiple imputation *F*-test procedure (see, e.g., van Buuren, 2018)

Sensitivity Analysis of Possible Threats to Valid Inferences

We conducted sensitivity tests to assess whether measurement non-invariance or item misfit may threaten valid inferences of population differences in children's overall development.

Measurement Non-Invariance. We conducted a sensitivity check to assess whether measurement non-invariance resulting from differential item functioning (DIF) threatens inferences about between-group differences in scores. We highlight here only the essential details of our procedure and refer the reader to the supplemental materials for technical details.

We first screened each item for DIF (uniform and non-uniform) across race, ethnicity, household income, caregiver's educational attainment, and enrollment in governmental services. For items exhibiting statistically significant DIF, we created a sequence of items ordered from smallest- to largest- *p*-value. Iteratively we removed the item in the sequence, refit the 2PL model in (1) and (2), extracted new EAP scores, refit Model 6. We recorded whether there was evidence that the coefficients differenced in significance or substantive size compared to the estimates where no items are removed.

In other words, given the KMT is a population measure, the presence of DIF is only concerning if it leads to different conclusions about group differences. It is well established in IRT literature that the presence of DIF is not sufficient to conclude that inferences are invalid (see, e.g., Chalmers, 2014).

Item Misfit. In addition to measurement non-invariance, we assessed whether item misfit resulting from a poorly fitting assumed item response function threatens valid inference about population differences. We proceeded in this evaluation in three phases. First, using the guidelines provided by Maydeu-Olivares (2014), we identified poorly fitting items as those with root mean square error (RMSEA) statistics greater than .08. For each of the identified items, we specified an item response model that employs a third-order monotonic polynomial to relax the traditional linearity assumption. We then compared scores from the model with monotonic polynomials specified for misfitting items to the original model in (1) and (2). Controlling for children's age, we considered part correlations less than .95 as evidence that conclusions risk being sensitive to item misfit.

Reliability and Errors of Measurement

We assessed the precision of scores in two ways. First, we calculate the marginal reliability statistic ($r_{XX'}$) proposed by Thissen and Wainer using the standard errors for the EAP estimates. Although the $r_{XX'}$ is a valid measure of reliability, as a marginal statistic it overestimates the reliability when a child's score is only to be compared to scores from their peers. To evaluate the precision of scores at conditional on a child's age, we fit a generalized additive model

for location scale and shape (GAMLSS; Rigby & Stasinopoulos, 2005) to estimate age-conditional variances in EAP scores. Using the within age variance of EAP scores and the conditional standard error of measurement (CSEM), we then calculated the expected conditional reliability value ($r_{XX'|\mathrm{AGE}_i}$) for each child,

$$r_{XX'|\mathrm{AGE}_i} = 1 - \frac{\widehat{\mathrm{CSEM}}_i^2}{\hat{\mathrm{Var}}\left\{\hat{\theta}^{\mathrm{EAP}}|\mathrm{AGE}_i\right\}}, \#(2)$$

where $\hat{\theta}^{\mathrm{EAP}}$ is the EAP score. We evaluated the average $r_{XX'|\mathrm{AGE}_i}$ values calculated in the previous step, in addition to evaluating the expected reliability $r_{XX'|\mathrm{AGE}_i}$ at each age. We used $r_{XX'|\mathrm{AGE}_i} = .80$ as a cutoff for the minimal reliability at each age. Such a reliability value may be below traditional guidelines for individual assessments (i.e., when the conclusions are drawn regarding individuals). However, population measures like the KMT are intended to produce statistics that aggregate across individual scores, thereby washing out measurement error across individual score.

Results

Table 1 reports the demographics of our Time 1 and Time 2 samples to those for the US population using nationally representative data from the NSCH provided by the Child and Adolescent Health and Measurement Initiative (2022).[1]

Table 1
Sample demographics and family characteristics.

	Time 1 Sample		Time 2 Sample		US Population[1]	
	N = 3,413	%	*N* = 70	%	N = 36.61 (x10^6)	%
	Child Characteristics (at Time 1)					
Sex						
Female	1595	46.7	40	57.1	5.75	48.9
Male	1818	53.3	30	42.9	6.02	51.1
Age						
0–11 mo.	554	16.2	17	24.3	1.85	15.7
12–23 mo.	567	16.6	21	30	1.94	16.5
24–35 mo.	645	18.9	22	31.4	1.99	16.9
36–47 mo.	639	18.7	10	14.3	2.02	17.1
48–59 mo.	616	18.1	-	-	2.04	17.3
60–71 mo.	392	11.5	-	-	1.94	16.5
Race/Ethnicity						
White, non-Hispanic	1911	55.9	41	58.6	6.11	51.9
Black, non-Hispanic	326	9.6	6	8.6	1.40	11.9
Other/Two or more, non-Hisp.	429	12.6	8	11.4	1.42	12.1
Hispanic	747	21.9	15	21.4	2.85	24.2
Adverse Experiences						
No ACEs	2504	73.3	57	81.4	22.4	66.1
1 ACE	474	13.9	12	17.1	6.57	19.4
2 + ACEs	436	12.8	1	1.5	4.92	14.5
Overall Health						
Excellent	1446	49.6	21	65.6	24.0	65.7
Very good	1074	36.8	9	28.1	8.91	24.4
Good	327	11.2	2	6.3	3.04	0.1
Fair or poor	70	2.4	0	0.0	0.56	< 0.1
	Caregiver/Household Characteristics (at Time 1)					
Educational Attainment						
Less than HS Diploma	153	4.5	0	0.0	4.60	12.7
HS Diploma	558	16.3	8	11.8	5.76	15.9
Some College or AA/AS	1192	34.9	11	16.2	10.20	28.0
BA/BS+	1510	44.2	49	72.0	15.80	43.5
Household Income[2]	3411	$68,700	70	$75,900	35.02	$69,700
Program Enrollment						
Food stamp or SNAP	653	19.1	11	16.2	35.41	20.3
Governmental healthcare	1582	46.4	21	30.9	33.24	19.0
WIC	929	27.2	16	23.5	35.32	20.2
Cash Assistance Program	193	5.7	1	1.5	35.32	20.2
Free and reduced price lunch	843	24.7	13	19.1	20.35	20.3
Anxiety and Depress. Sxs (0-Not at all, 5-Everyday)						

	Time 1 Sample		Time 2 Sample		US Population[1]
Feeling nervous, anxious, or on edge	2909	M = .75	36	M = .61	
Feeling down, depressed not being able to stop or control worrying hopeless	2899	M = .47	36	M = .22	
Not being able to stop or control worrying	2910	M = .60	36	M = .31	
little interest or pleasure in doing things	2911	M = .50	36	M = .25	

Time 1 Sample (N = 3,413)

At Time 1, parents with male children (53.3% vs. 51.1% nationally) were slightly more likely to respond than parents with female children. Compared to national estimates, parents in our sample were more likely to report their child had an ACE count equal to zero (73.3% vs. 66.1% nationally), but these parents also were less likely to report that their child was in excellent overall health (49.6% vs. 65.7% nationally). Our sample was slightly more likely to identify as White, non-Hispanic (55.9% vs. 51.9% nationally), and slightly less likely to identify as Black, non-Hispanic (9.6% vs. 11.9% nationally) or Hispanic (21.9% vs. 24.2% nationally). Parents in our sample were also slightly more likely to have attained a BA/BS (44.2% vs. 43.5% nationally), and less likely to report having not attained a high school diploma or GED (4.5% vs. 12.7% nationally). The median income in our sample was $68,700 (in 2022 USD), which effectively matches the national estimate of $69,700. Compared to national estimates, our sample was more likely to be enrolled in (1) a government healthcare program (46.4% vs. 19.0% nationally), (2) WIC (27.2% VS. 20.2% nationally), or (3) free and reduced-price lunch (24.7% vs. 20.3% nationally). However, parents in our sample were less likely to report enrollment in a cash assistance program (5.7% vs. 20.2% nationally) or receive food stamps/SNAP (19.1% vs. 20.3% nationally).

Time 2 Sample (N = 70)

We obtained follow-up data from more female children than male children at Time 2 (57.1% vs. 48.9% nationally). This sample also tended to be from families with higher levels of income ($75,900 vs. $69,700 nationally), higher levels of educational attainment as indicated by attainment of a BA/BS degree (72.0% vs. 43.5% nationally), fewer number of ACEs (No ACEs: 81.4% vs. 66.1% nationally; 1 ACE: 17.1% vs. 19.4%; 2 + ACEs: 1.5% vs. 14.5% nationally). However, Time 2 parents reported similar levels of children in either excellent (65.6% vs. 65.7% nationally) or good (28.1% vs. 24.4% nationally) overall health compared to the U.S. population.

Evidence Based on Test Content

The proportion of items categorized as motor or physical development, cognition or language development, and social/emotional development is presented in Fig. 1 below. Motor and physical development represented most items (55%) assigned in the first year of life, but these items represent only 19% of administered items for children aged 2 years and older. In contrast, items identified as measuring cognitive or language development represent less than a quarter (24%) of items assigned to newborns, but these items represent a majority beginning at age 3 (Max = 57%). The percentage of items reflecting social/emotional development varies between 22% (for newborns) and 31% (for two-year-olds).

Evidence Based on Other Variables

Convergent Validity Evidence

Part correlations of Kidsights scores with Bayley-4 and the WJ-ECAD scores administered concurrently (i.e., each administered at Time 2) are presented in Tables 2 & 3, respectively. With the Bayley-4, correlation coefficients were greater or equal to than the minimally acceptable threshold or r = .50 (see Table 2). Kidsights scores at Time 2 were correlated with Bayley-4 Expressive Communication scores (r = .63, p < .001) and correlated with Baley-4 Fine Motor scores (r = .50, p < .001). For the WJ IV ECAD, the correlation between the Kidsights scores at Time 2 and the WJ IV ECAD scores for General Intellectual Ability- Early Development scores, Sentence Repetition scores, Expressive Language, Verbal Analogies, and Rapid Picture Naming test activities all met the minimally acceptable threshold (see Table 3; min: r = .54, p < .001, max: r = .66, p < .001). However, the part correlation between Kidsights scores at Time 2 with the WJ IV ECAD Picture Vocabulary scores (r = .46, p < .01; r = .48, p<.001) were below the minimally acceptable threshold. Kidsights scores were not significantly correlated with Memory for Names, Sound Blending and Visual Closure scores from the WJ IV ECAD.

Table 2
Part correlations with scores from Bayley-4 administered at Time 2 (n = 37).

	(1)	(2)	(3)	(4)	(5)	(6)
1. Time 1 Kidsights	1.00					
2. Time 2 Kidsights	.66***	1.00				
	(.10)					
3. Cognition	.74***	.59***	1.00			
	(.07)	(.12)				
4. Receptive Communication	.64***	.56***	.95***	1.00		
	(.09)	(.11)	(.01)			
5. Expressive Communication	.77***	.63***	.92***	.90***	1.00	
	(.06)	(.10)	(.02)	(.02)		
6. Fine Motor	.73***	.50***	.94***	.90***	.90***	1.00
	(.08)	(.13)	(.01)	(.02)	(.02)	
7. Gross Motor	.75***	.57***	.87***	.79***	.83***	.87***
	(.07)	(.12)	(.03)	(.05)	(.04)	(.03)
Note: *** $p < .001$						

Table 3
Part correlations with scores from WJ IV ECAD administered at Time 2 (n = 33).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1. Time 1 Kidsights	1.00									
2. Time 2 Kidsights	.86***	1.00								
	(.04)									
3. GIA - Early Development	.66***	.70***	1.00							
	(.09)	(.08)								
4. Expressive Language	.59***	.66***	.92***	1.00						
	(.11)	(.09)	(.02)							
5. Memory for Names	.28	.24	.34*	.30	1.00					
	(.23)	(.22)	(.15)	(.16)						
6. Sound Blending	.37	.22	.38**	.17	− .02	1.00				
	(.21)	(.21)	(.14)	(.16)	(.18)					
7. Picture Vocabulary	.46**	.48***	.75***	.87***	.20	.08	1.00			
	(.16)	(.14)	(.04)	(.01)	(.16)	(.17)				
8. Verbal Analogies	.56***	.54***	.81***	.68***	− .01	.30	.57***	1.00		
	(.13)	(.12)	(.05)	(.08)	(.17)	(.16)	(.10)			
9. Visual Closure	− .08	.03	.24	.21	.28	− .15	.27	.06		
	(.33)	(.32)	(.22)	(.23)	(.20)	(.19)	(.21)	(.21)		
10. Sentence Repetition	.61***	.70*	.92***	.97***	.32*	.19	.73***	.67***	.14	
	(.11)	(.08)	(.02)	(.00)	(.16)	(.17)	(.02)	(.08)	(.23)	
11. Rapid Picture Naming	.54***	.63***	.82***	.68***	.29	.25	.42***	.60***	.05	.75***
	(.14)	(.11)	(.05)	(.08)	(.16)	(.17)	(.11)	(.10)	(.22)	(.06)
Notes: ***$p < .001$, **$p < .01$, *$p < .05$										

Table 4 provides convergent validity with scores derived from previously validated caregiver measures. Kidsights scores were highly correlated with D-scores (r = .93, p < .001) and CREDI-LF Overall scores (r = .89, p < .001). Correlation coefficients were also greater than the ideal threshold value (i.e., $r \geq .70$) for all CREDI subscores (min: r = .05, p < .001, max: r = .82, p < .001), ECDI2030 scores (r = .75, p < .001), and NOM early learning factor scores (r = .77, p < .001). Correlation coefficients were found to be greater than the minimally acceptable threshold value of r = .50 for HRTL physical development factor scores (r = .62, p < .001). Only the correlation with HRTL social/emotional factor scores (r = .39, p < .001) were found to below the minimum threshold value we set.

Table 4
Part correlations with CREDI and HRTL instruments administered at Time 1 (N = 3,413).

			(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(1
	1.	Kidsights	1.00											
	2.	D-scores	.93***	1.00										
			(.01)											
CREDI	3.	Overall	.89***	.91***	1.00									
			(.01)	(.01)										
	4.	Motor	.82***	.86***	.87***	1.00								
			(.01)	(.01)	(.01)									
	5.	Cognition	.86***	.86***	.92***	.91***	1.00							
			(.01)	(.01)	(.01)	(.01)								
	6.	Language	.85***	.87***	.94***	.78***	.87***	1.00						
			(.01)	(.01)	(.01)	(.01)	(.01)							
	7.	Soc./Emot.	.82***	.81***	.87***	.85***	.97***	.80***	1.00					
			(.01)	(.01)	(.01)	(.01)	(.01)	(.01)						
	8.	ECDI2030	.75***	.70***	.71***	.67***	.71***	.76***	.66***	1.00				
			(.01)	(.02)	(.02)	(.02)	(.02)	(.02)	(.02)					
HRTL	9.	Physical	.62***	.52***	.45***	.49***	.44***	.43***	.41***	.48***	1.00			
			(.02)	(.03)	(.05)	(.05)	(.04)	(.03)	(.04)	(.02)				
	10.	Early Learning	.77***	.63***	.62***	.56***	.62***	.59***	.59***	.56***	.43***	1.00		
			(.01)	(.02)	(.03)	(.03)	(.03)	(.02)	(.03)	(.02)	(.02)			
	11.	Soc./Emot.	.39***	.37***	.42***	.39***	.39***	.34***	.38***	.30***	.16***	.21***	1.00	
			(.02)	(.03)	(.04)	(.04)	(.04)	(.04)	(.04)	(.03)	(.03)	(.03)		
	12.	Self Reg.	.38***	.33***	.39***	.33***	.37***	.34***	.35***	.24***	.16***	.27***	.36***	1.
			(.02)	(.03)	(.04)	(.04)	(.04)	(.03)	(.04)	(.03)	(.03)	(.03)	(.02)	
	13.	GSED PF	− .20***	− .17***	− .14***	− .15***	− .18***	− .16***	− .18***	− .16***	− .12***	− .14***	− .43***	− .
			(.02)	(.02)	(.02)	(.02)	(.02)	(.02)	(.02)	(.02)	(.02)	(.02)	(.02)	(.

Notes: ***p < .001, **p < .01, *p < .05

Discriminant Validity Evidence

As can be seen in Table 4, part correlations of Kidsights scores with HRTL self-regulation scores (r = .38, p < .001) and GSED-PF psychosocial problems scores (r = − .20, p < .001) provide supportive evidence of discriminant validity. The magnitude of each correlation was less than the magnitudes of the correlations used as evidence of convergent validity evidence. These correlations were also below the maximum threshold value ($|r|$ < .50) and in the expected directions (i.e., a positive correlation with HRTL self-regulation scores and a negative correlation with GSED-PF psychosocial problem scores).

Predictive Validity Evidence

Correlations with Kidsights scores at Time 1 and Bayley-4 or WJ IV ECAD scores obtained 12–18 months later at Time 2 provide predictive validity evidence (see Tables 2 & 3). Kidsights scores at Time 1 were found to predict all Bayley-4 scores at Time 2; each correlation coefficient was positive, significant, and substantively large (see Table 2; min: r = .64, p < .001; max: r = .77, p < .001). Similarly, Kidsights scores at Time 2 predicted WJ IV ECAD (see Table 3) GIA Early Development scores (r = .70, p < .001), Sentence Repetition scores (r = .70, p = .04), Expressive Language scores (r = .66, p < .001), Verbal Analogies scores (r = .54, p < .001), and Picture Vocabulary scores (r = .48, p < .001). However, we failed to find that Kidsights scores at Time 1 predicted WJ IV Memory for Names scores, Sound Blending scores, or Visual Closure scores.

Criterion Associations

We find favorable evidence that the associations of between Kidsights scores and criterion variables are consistent with those expected from theory. Table 5 presents the coefficient estimates for Models 1–6. Females (Model 1) and children reported in very good or excellent health (Model 2) exhibit higher average scores than their peers (Model 1: FEMALE: Est. = .149, SE = 0.035; $p < .001$, ES = .155; Model 2: HEALTHY 2: Est. = 0.501, SE = 0.052, $p < .001$, ES = 0.522). Simultaneous testing (Model 3 vs. Model 2: $F = 12.397$, $df_1 = 3$, $df_2 = 3387.414$) and pairwise tests indicate that children from parents with at college attainment exhibit high average scores than those who only possess a high school diploma (Model 3: SOMECOLL: Est. = 0.147, SE = 0.049; $p = .003$, ES = .153; BS: Est. = 0.182, SE = 0.051; $p < .001$, ES = 0.190), but no significant differences were found between parents who had not attained a high school diploma versus those who had a high school diploma ($p = .575$). Enrollment in governmental services (e.g., SNAP, WIC, etc.) was associated with lower average scores, even after controlling for caregiver's educational attainment (Model 4: GOVT: Est. = -0.122, SE = 0.04, $p = .001$, ES = -0.127). As expected, after controlling for caregiver's educational attainment and enrollment in governmental programs and services, we simultaneous testing indicated between-group differences in average scores race and ethnicity (Model 5 vs. Model 4, $F = 3.101$, $df_1 = 3$, $df_2 = 3379.131$, $p = .026$), but pairwise t-tests were each not significant (Model 5: BLACK: $p = .442$; HISP: $p = .068$; WHITE: $p = .880$). Lastly, caregiver depression and anxiety symptoms and the child's adverse experience were also negatively associated with scores (Model 6 vs. Model 5: $F = 10.459$, $df_1 = 2$, $df_2 = 2323.525$, $p < .001$; Model 6: DEPANX: Est. = -0.059, SE = 0.017, $p < .001$, ES = -0.069, ACEs: Est. = -0.038, SE = 0.018, $p = 0.038$; ES = -0.040).

Table 5

Validity evidence from criterion associations at Time 1 (N = 3,413).

	M1	M2	M3	M4	M5	M6
Female	0.149***	0.161***	0.161***	0.166***	0.167***	0.209***
	(0.035)	(0.035)	(0.034)	(0.034)	(0.034)	(0.036)
	[0.155]	[0.168]	[0.168]	[0.173]	[0.174]	[0.217]
Child in Very Good or Excellent Overall Health		0.501***	0.487***	0.462***	0.457***	0.431***
		(0.052)	(0.052)	(0.053)	(0.053)	(0.053)
		[0.522]	[0.507]	[0.480]	[0.476]	[0.449]
No High School Diploma			-0.049	-0.046	-0.041	-0.054
			(0.086)	(0.086)	(0.086)	(0.086)
			[-0.051]	[-0.048]	[-0.043]	[-0.056]
Some College or Associate's Degree			0.170***	0.147**	0.137**	0.152**
			(0.048)	(0.048)	(0.049)	(0.049)
			[0.177]	[0.153]	[0.143]	[0.158]
Bachelor's Degree or Higher			0.250***	0.182***	0.169**	0.182***
			(0.047)	(0.051)	(0.052)	(0.052)
			[0.261]	[0.190]	[0.176]	[0.189]
Government Program or Services				-0.122**	-0.111**	-0.073
				(0.038)	(0.039)	(0.040)
				[-0.127]	[-0.116]	[-0.075]
Black					0.054	0.061
					(0.070)	(0.070)
					[0.056]	[0.063]
Hispanic					-0.104	-0.098
					(0.057)	(0.057)
					[-0.108]	[-0.102]
White					0.008	0.004
					(0.052)	(0.052)
					[0.008]	[0.004]
Caregiver Depression and Anxiety Symptoms						-0.059***
						(0.017)
						[-0.069]
Child's ACEs						-0.038*
						(0.018)
						[-0.040]
Notes: ***p < .001, **p < .01, *p < .05.						

Sensitivity Analysis of Possible Threats to Valid Inference

Measurement Non-Invariance

In total, we found approximately one-third of all items in the Kidsights Measurement Tool item bank exhibited uniform or non-uniform DIF (see Supplemental Table 2). As detailed in the supplementary materials, we failed to find that the presence of DIF threatens inferences about between-group differences in young children's development. Specifically, criterion associations retain their substantive size, direction, and significance even after removing all items exhibiting DIF (see Supplemental Fig. 1).

Item Misfit

In total, we found that nearly one-fifth of items exhibited RMSEA values greater than .08, suggesting a poorly fitting item response model for a substantial number of items (see Supplemental Table 2). As shown in the supplemental materials, scores remain correlated at r = .99 (see Supplemental Fig. 2) after relaxing the linearity assumption by specifying a monotonic polynomial item response function. Thus, we do not find evidence that inferences are threatened from item misfit.

Reliability and Errors of Measurement

Reliability and errors of measurement estimate provide supportive evidence that measurement error does not unduly threaten population-level inferences. As expected, the marginal reliability was found to be very strong, $r_{XX'}$ = .99. Controlling for variation due to children's age, we found an average conditional reliability estimate of $\bar{r}_{XX'|\mathrm{AGE}}$= .92. Moreover, the expected conditional reliability was found to be greater than the minimum threshold value of .80 across all ages (see Fig. 3).

[1] The NSCH is a nationally representative household survey of the US population directed by the Maternal and Child Health Bureau (MCHB) of the U.S. Department of Health and Human Services. The goal of the NSCH is to produce national and state level estimates of children's health and well-being of children and their families (MCHB, 2023). In generating U.S. population estimates for comparison to our sample, we pooled all household data with a child 0-5 years provided by the Child and Adolescent Health and Measurement Initiative (2022) collected between 2016-2020.

Discussion

Population-based measurement of early child development has the potential to generate new insight into the emergence of disparities in the earliest years of life. Because tracking and addressing early disparities is a cornerstone of effective public health systems, such data are critical. However, documenting these disparities requires feasible and reliable measurement tools that can be used across populations, and at present, governments have limited means to examine trends in child development. Our study documents multiple types of validity for the KMT, demonstrating its feasibility for use as a population-level measure of child development. We demonstrate content validity evidence of the KMT, as well as convergent and discriminant evidence from concurrently administered instruments, predictive validity with direct-observation instruments, criterion validity with known correlates of children's development, and sufficient reliability for the purposes of drawing conclusions about groups of children. In sum, this study offers evidence that the Kidsights Measurement Tool meets the psychometric requirements detailed by the APA *Standards* for its intended use as a population-based measurement tool appropriate for children birth to age 5 years. Below we outline our findings in greater depth.

First, we found that KMT test content demonstrates concordance with key domains of child development including cognitive, language, motor and social/emotional development. Although the proportion of items representing different domains varies by age, overall, the balance of domains represents relevant domains at each age, which was our goal. The process of selecting items was aided considerably by the reliance on existing scales of child development including the CREDI, GSED, and HRTL, each of which was developed through careful identification of developmental milestones in early childhood (e.g., McCoy et al, 2018; Cavallera et al., 2023; Ghandour et al., 2019).

Second, we found that Kidsights scores showed expected predictive and concurrent associations with scores from both direct observations and other parent-report measures designed for use at the population level, the CREDI, HRTL and ECDI. Beginning with the sample of 70 children assessed at two time points, we found evidence of predictive validity. Kidsights scores strongly predicted all Bayley-4 subscores observed 12–18 months later, meeting or exceeding the minimum acceptable threshold for association. Kidsights scores were also related to several of the WJ IV ECAD scores obtained 12–18 months after Kidsights scores were collected. We found similar evidence of convergent validity when the KMT was administered concurrently with the Bayley-4 and the WJ IV ECAD, as well as associations with previously validated parent reported measures (i.e., GSED-SF, CREDI, and HRTL). While associations were largely confirmed by our data, we did not find that all subscales of the WJ IV ECAD were associated with Kidsights scores from a concurrent administration of the KMT. Specifically, the WJ IV ECAD subscales of memory for names, sound blending, and visual closure did not show concurrent associations with KMT scores, while picture vocabulary, verbal analogies, sentence repetition and rapid picture naming did. While all of these subscales are intended to index general intellectual ability (Shrank et al., 2018), it is not clear from our study why some were related to KMT and some were not, although some of the subscales were very difficult for children in our sample so the lack of positive association may have been due to overall low scores. Future work with KMT and indices of academic achievement, including following children later into the primary school years, may help clarify these findings. At the same time, the KMT is intended to be a holistic measure of child development rather than indexing specific skills, and thus may show greater sensitivity to some aspects of later academic achievement than others.

Third, Kidsights scores met acceptable standards of measurement invariance and score precision. We found scores to be sufficiently precise (i.e., reliability) to support inferences of group mean differences. Through sensitivity analyses, we found no evidence that (a) assumption violations in modeling the internal structure (i.e., item misfit), nor (b) measurement non-invariance indicated by DIF threatened inferences about population mean differences.

Fourth, in assessing criterion validity, we found negative associations between caregiver enrollment into governmental services and programs, caregiver depressive and anxiety symptoms, and the child's count of ACEs. Children whose parents reported enrollment in government services (indicating household economic stress), greater depressive and anxious symptomology, and more ACEs had lower KMT scores, all of which is consistent with existing research on threats to early child development (Petterson et al., 2001; Treat, et al., 2020).

Our results also identify areas for further inquiry. We found smaller concurrent associations between KMT scores and HRTL subscales focused on social/emotional and self-regulatory development than with the HRTL early learning and physical development subscales. While the Kidsights Measurement

Tool contains indicators of social/emotional development, the majority of items are focused on cognitive, language and physical development. Further, HRTL social/emotional and self-regulatory subscales include indicators of problem behaviors as well as normative development, whereas the Kidsights Measurement Tool was focused exclusively on indexing social/emotional and regulatory competence. Going forward, we will continue to test Kidsights Measurement Tool's sensitivity to early social and emotional development, including through the development of a new scale focused specifically on manifestations of early psychosocial stress (author redacted, under review).

Conclusions

Taken together, results suggest that Kidsights is a reliable and valid approach to measurement of holistic child development at the population level for children birth to age 5. Identifying and tracking early disparities has significance for ensuring lifelong health and wellbeing, particularly for populations of young children who are vulnerable to high levels of stress due to the quality of their environments. The KMT tool opens the door to more extensive population-level monitoring of young children's development beginning in the earliest years of life, when addressing disparities is most cost effective. While we acknowledge that representative sampling is needed to ensure accurate population-level monitoring, our tool represents a critical step forward in generating evidence of early disparities.

Declarations

Ethical Approval

All protocols were reviewed by the Institutional Review Board of University of Nebraska Medical Center and the World Health Organization's Ethics Review Committee (ID ERC.0003514). All participants provided documentation acknowledging informed consent.

Competing Interests

We have no known competing or conflicts of interest to disclose.

Authors' Contributions

M.W. oversaw all psychometric validity and reliability analysis, generated all tables and figures, and led the writing of the methods and results. K.H. administered direct-observation instruments, completed concurrent and predictive validity evidence, and assisted with the writing of the main text. J.J., K.T., K.J., Y.E.G., L.F., and A.S. assisted with recruitment and data collection. A.R. led the writing of the background, discussion, and conclusion. All authors reviewed the manuscript.

Funding

Support for this research was provided by the Buffett Early Childhood Fund, Imaginable Futures, Overdeck Family Foundation, Pritzker Children's Initiative, and Valhalla Foundation.

Availability of Data and Materials

The data that support the findings of this study are available from the corresponding author, M.W., upon reasonable request.

References

1. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
2. Bassok, D., & Latham, S. (2017). Kids today: The rise in children's academic skills at kindergarten entry. *Educational Researcher*, *46*(1), 7-20.
3. Burchinal, M., McCartney, K., Steinberg, L., Crosnoe, R., Friedman, S. L., McLoyd, V., ... & NICHD Early Child Care Research Network. (2011). Examining the Black–White achievement gap among low-income children using the NICHD study of early child care and youth development. *Child development*, *82*(5), 1404-1420.
4. Centers for Disease Control (2023). Healthy People 2030: Objectives and data. Retrieved from https://health.gov/healthypeople/objectives-and-data/browse-objectives/children/increase-proportion-children-who-are-developmentally-ready-school-emc-d01
5. Child and Adolescent Health Measurement Initiative (CAHMI) (2022). 2016-2020 National Survey of Children's Health, SPSS Indicator dataset. Data Resource Center for Child and Adolescent Health supported by Cooperative Agreement U59MC27866 from the U.S. Department of Health and Human Services, Health Resources and Services Administration (HRSA), Maternal and Child Health Bureau (MCHB). Retrieved 08/31/2022 from childhealthdata.org.
6. Cappa C, Petrowski N, De Castro EF, Geisen E, LeBaron P, Allen-Leigh B, Place JM, Scanlon PJ. Identifying and Minimizing Errors in the Measurement of Early Childhood Development: Lessons Learned from the Cognitive Testing of the ECDI2030. *International Journal of Environmental Research and Public Health*. 2021; 18(22):12181. https://doi.org/10.3390/ijerph182212181References
7. Campbell, F., Conti, G., Heckman, J. J., Moon, S. H., Pinto, R., Pungello, E., & Pan, Y. (2014). Early childhood investments substantially boost adult health. *Science*, *343*(6178), 1478-1485.
8. Cavallera, V., Lancaster, G., Gladstone, M., Black, M. M., McCray, G., Nizar, A., ... & Janus, M. (2023). Protocol for validation of the Global Scales for Early Development (GSED) for children under 3 years of age in seven countries. *BMJ open*, *13*(1), e062562.

9. Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of statistical Software*, *48*, 1-29.
10. Feder, A., Fred-Torres, S., Southwick, S. M., & Charney, D. S. (2019). The biology of human resilience: opportunities for enhancing resilience across the life span. *Biological psychiatry*, *86*(6), 443-453.
11. Friedman-Krauss, A., & Barnett, S. (2020). Access to high-quality early education and racial equity. *National Institute for Early Education Research. Retrieved from https://nieer. org/wp-content/uploads/2021/02/Special-Report-Access-to-High-Quality-Early-Education-and-Racial-Equity. pdf*.
12. Garner, A., Yogman, M., & Committee on Psychosocial Aspects of Child and Family Health. (2021). Preventing childhood toxic stress: partnering with families and communities to promote relational health. *Pediatrics*, *148*(2).
13. Ghandour, R. M., Moore, K. A., Murphy, K., Bethell, C., Jones, J. R., Harwood, R., ... & Lu, M. (2019). School readiness among US children: Development of a pilot measure. *Child Indicators Research*, *12*, 1389-1411.
14. Halle, T., Forry, N., Hair, E., Perper, K., Wandner, L., Wessel, J., & Vick, J. (2009). Disparities in early learning and development: lessons from the Early Childhood Longitudinal Study–Birth Cohort (ECLS-B). *Washington, DC: Child Trends*, 1-7.
15. Halpin, P., de Castro, E. F., Petrowski, N., & Cappa, C. (2023). Monitoring Early Childhood Development at the Population Level: The ECDI2030. Retrieved from
16. Gluckman, P. D., Hanson, M. A., Wintour, E. M., & Owens, J. A. (2006). Early life origins of health and disease.
17. Gollenberg, A. L., Lynch, C. D., Jackson, L. W., McGuinness, B. M., & Msall, M. E. (2010). Concurrent validity of the parent-completed Ages and Stages Questionnaires, with the Bayley Scales of Infant Development II in a low-risk sample. *Child: care, health and development*, *36*(4), 485-490.
18. Heckman, J. J. (2006). Skill formation and the economics of investing in disadvantaged children. *Science*, *312*(5782), 1900-1902.
19. Janus, M., Brinkman, S. A., & Duku, E. K. (2011). Validity and psychometric properties of the early development instrument in Canada, Australia, United States, and Jamaica. *Social Indicators Research*, *103*, 283-297.
20. Jeong, J., Franchett, E. E., Ramos de Oliveira, C. V., Rehmani, K., & Yousafzai, A. K. (2021). Parenting interventions to promote early child development in the first three years of life: A global systematic review and meta-analysis. *PLoS medicine*, *18*(5), e1003602.
21. Klein-Radukic, S., & Zmyj, N. (2023). The predictive value of the cognitive scale of the Bayley Scales of Infant and Toddler Development-III. *Cognitive Development*, *65*, 101291.
22. Kuhfeld, M., Soland, J., Pitts, C., & Burchinal, M. (2020). Trends in children's academic skills at school entry: 2010 to 2017. *Educational Researcher*, *49*(6), 403-414.
23. Lawlor, J., Thomas, C., Guhin, A. T., Kenyon, K., Lerner, M. D., & Drahota, A. (2021). Suspicious and fraudulent online survey participation: Introducing the REAL framework. Methodological Innovations, 14(3). https://doi.org/10.1177/20597991211050467
24. Maydeu-Olivares A., Joe H. (2014). Assessing approximate fit in categorical data analysis. Multivariate Behavioral Research, 49, 305-328. doi: 10.1080/00273171.2014.911075
25. McCray, G., McCoy, D., Kariger, P., Janus, M., Black, M. M., Chang, S. M., ... & Gladstone, M. (2023). The creation of the Global Scales for Early Development (GSED) for children aged 0–3 years: combining subject matter expert judgements with big data. *BMJ Global Health*, *8*(1), e009827.
26. McCoy, D. C., Waldman, M., Team, C. F., & Fink, G. (2018). Measuring early childhood development at a global scale: Evidence from the Caregiver-Reported Early Development Instruments. *Early childhood research quarterly*, *45*, 58-68.
27. Molnár, G., & Hermann, Z. (2023). Short-and long-term effects of COVID-related kindergarten and school closures on first-to eighth-grade students' school readiness skills and mathematics, reading and science learning. *Learning and Instruction*, *83*, 101706.
28. Paschall, K., Anderson Moore, K., Pina, G., & Anderson, S. (2020). Comparing the national outcome measure of healthy and ready to learn with other well-being and school readiness measures. *Child Trends*, 1-19.
29. Petterson, S. M., & Albers, A. B. (2001). Effects of poverty and maternal depression on early child development. *Child development*, *72*(6), 1794-1813.
30. R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
31. Reynolds, A. J., Temple, J. A., Ou, S. R., Robertson, D. L., Mersky, J. P., Topitzes, J. W., & Niles, M. D. (2007). Effects of a school-based, early childhood intervention on adult health and well-being: A 19-year follow-up of low-income families. *Archives of pediatrics & adolescent medicine*, *161*(8), 730-739.
32. Rigby, R. A. and Stasinopoulos D. M. (2005). Generalized additive models for location, scale and shape,(with discussion), *Appl. Statist.*, **54**, part 3, pp 507-554
33. Rosseel Y (2012). "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, **48**(2), 1–36. doi:10.18637/jss.v048.i02.
34. Ryberg, R., Wiggins, L., Moore, K. A., Daily, S., Piña, G., & Klin, A. (2022). Measuring state-level infant and toddler well-being in the United States: Gaps in data lead to gaps in understanding. *Child indicators research*, *15*(3), 1063-1102.
35. Schafer, J. L., & Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological methods*, *7*(2), 147–177.
36. Schrank, F. A., Wendling, B. J., Flanagan, D. P., & McDonough, E. M. (2018). The Woodcock–Johnson IV Tests of Early Cognitive and Academic Development. *Contemporary intellectual assessment: Theories, tests, and issues*, 283-301.
37. Seiden, J., Waldman, M., McCoy, D. C., Fink, G. (2021) CREDI: data management & scoring manual. https://credi.gse.harvard.edu/credi-materials.
38. Shonkoff, J. P., Garner, A. S., Committee on Psychosocial Aspects of Child and Family Health, Committee on Early Childhood, Adoption, and Dependent Care, and Section on Developmental and Behavioral Pediatrics, Siegel, B. S., Dobbins, M. I., Earls, M. F., ... & Wood, D. L. (2012). The lifelong effects of early childhood adversity and toxic stress. *Pediatrics*, *129*(1), e232-e246.

39. Shonkoff, J. P., Slopen, N., & Williams, D. R. (2021). Early childhood adversity, toxic stress, and the impacts of racism on the foundations of health. *Annual Review of Public Health*, *42*, 115-134.
40. Treat, A. E., Sheffield-Morris, A., Williamson, A. C., & Hays-Grudo, J. (2020). Adverse childhood experiences and young children's social and emotional development: the role of maternal depression, self-efficacy, and social support. *Early Child Development and Care*, *190*(15), 2422-2436.
41. Waggoner, Philip D., Ryan Kennedy, and Scott Clifford, (2019). Detecting Fraud in Online Surveys by Tracing, Scoring, and Visualizing IP Addresses. Journal of Open Source Software, 4(37), 1285, https://doi.org/10.21105/joss.01285.
42. Walker, B. H., Brown, D. C., Walker, C. S., Stubbs-Richardson, M., Oliveros, A. D., & Buttross, S. (2022). Childhood adversity associated with poorer health: evidence from the US National Survey of Children's Health. *Child Abuse & Neglect*, *134*, 105871.
43. Weber, A. M., Rubio-Codina, M., Walker, S. P., Van Buuren, S., Eekhout, I., Grantham-McGregor, S. M., ... & Black, M. M. (2019). The D-score: a metric for interpreting the early development of infants and toddlers across global settings. *BMJ global health*, *4*(6), e001724.
44. Thissen, D. and Wainer, H. (2001). Test Scoring. Lawrence Erlbaum Associates.
45. UNICEF. (2023). Early Childhood Development Index 2030: Resources. https://data.unicef.org/resources/early-childhood-development-index-2030-ecdi2030/.
46. Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.
47. Van Buuren, S., Groothuis-Oudshoorn, K., Robitzsch, A., Vink, G., Doove, L., & Jolani, S. (2015). Package 'mice'. *Computer software*.
48. Veldhuizen, S., Clinton, J., Rodriguez, C., Wade, T. J., & Cairney, J. (2015). Concurrent validity of the Ages and Stages Questionnaires and Bayley Developmental Scales in a general population sample. *Academic pediatrics*, *15*(2), 231-237.
49. Voigt, R. G., Llorente, A. M., Jensen, C. L., Fraley, J. K., Barbaresi, W. J., & Heird, W. C. (2007). Comparison of the validity of direct pediatric developmental evaluation versus developmental screening by parent report. *Clinical pediatrics*, *46*(6), 523-529.
50. Yoshikawa, H., Wuermli, A. J., Raikes, A., Kim, S., & Kabay, S. B. (2018). Toward high-quality early childhood development programs and policies at national scale: Directions for research in global contexts. *Social Policy Report*, *31*(1), 1-36.

Figures

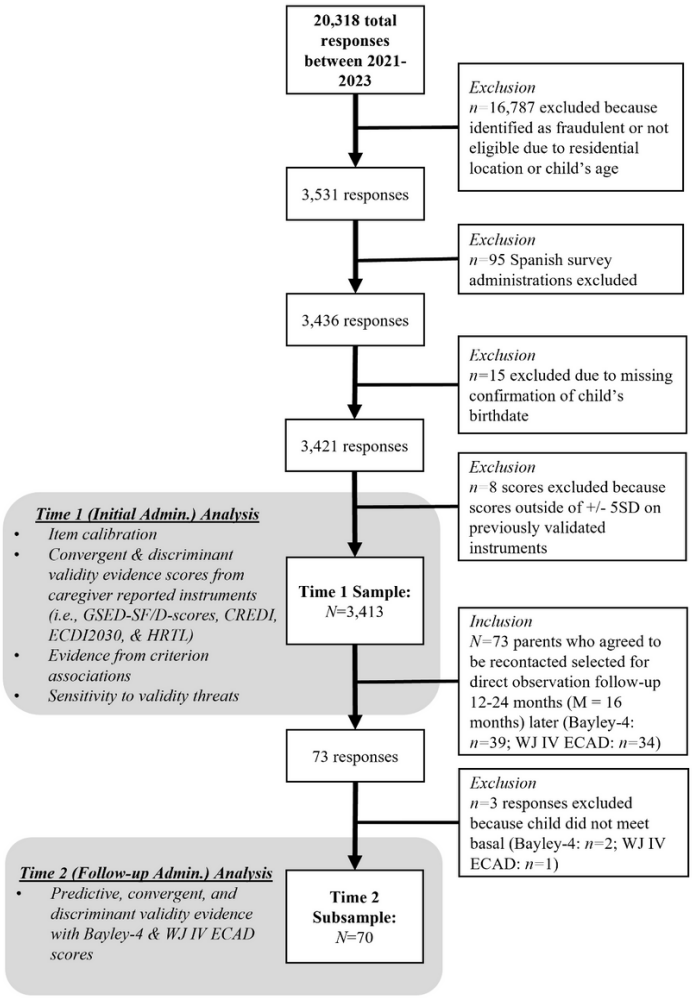

Figure 1

Flowchart of study selection based on inclusion and exclusion criteria

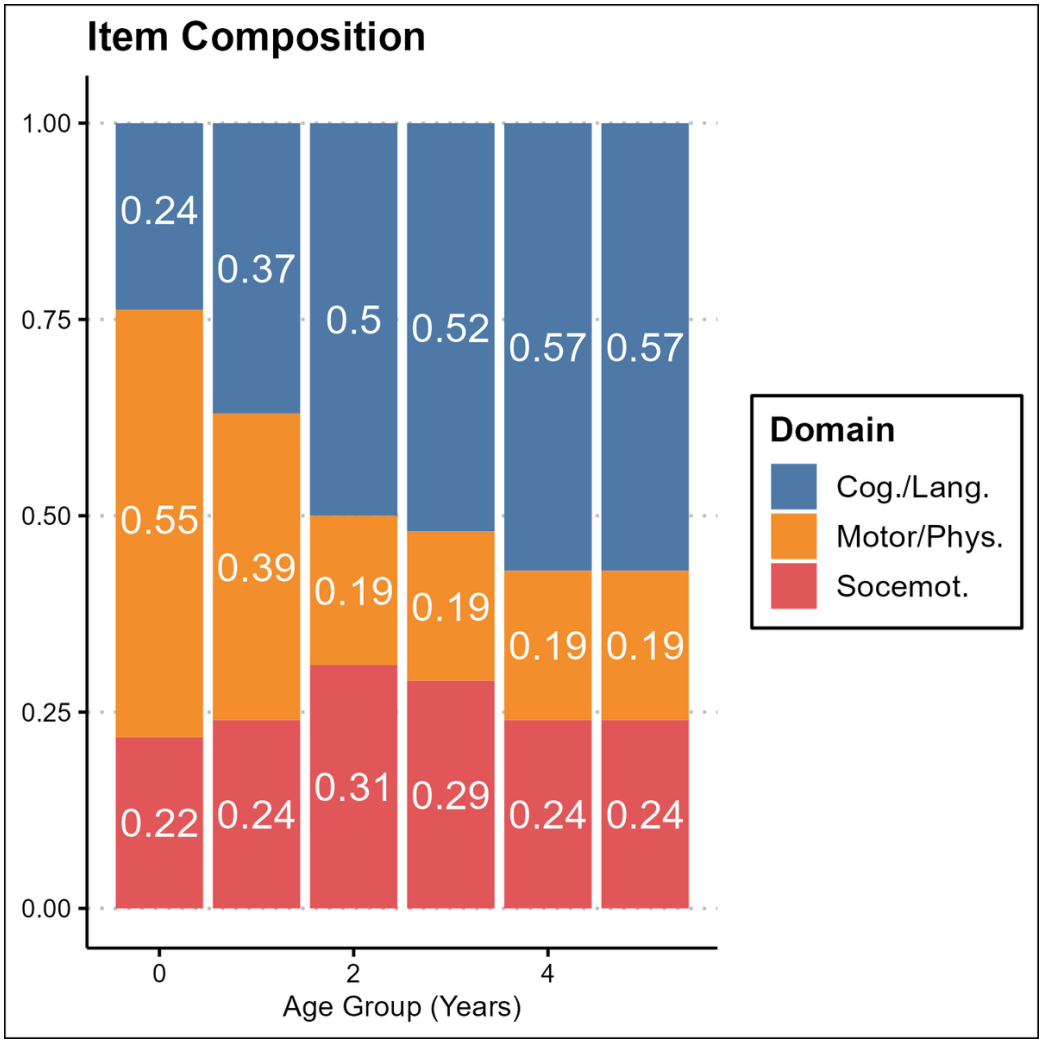

Figure 2

Kidsights Measurement Tool domain representation by child's age.

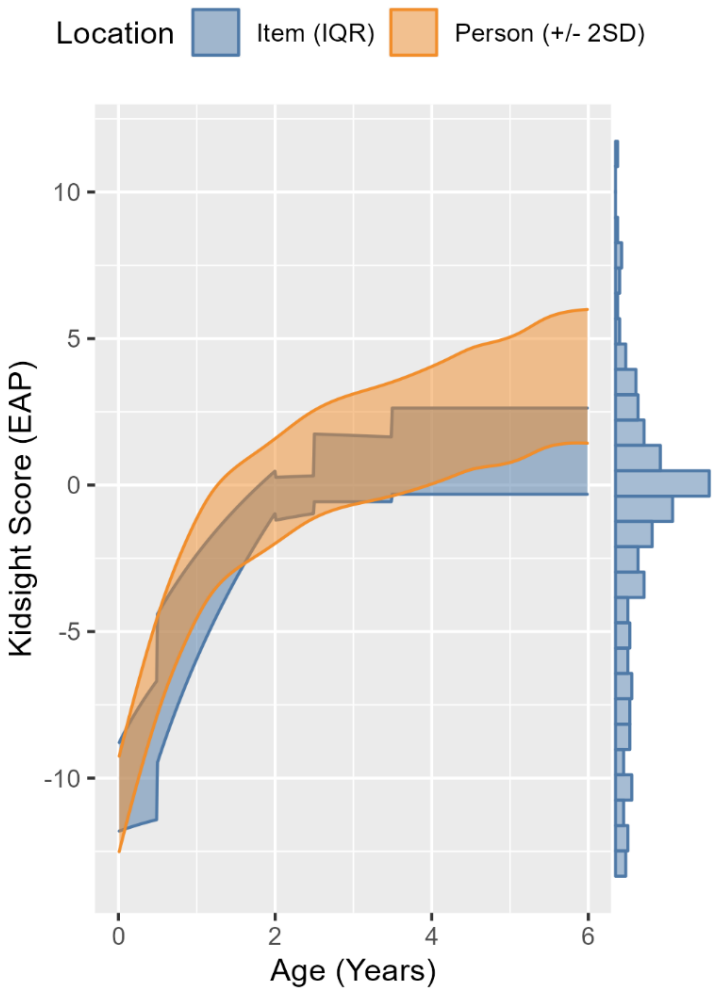

Figure 3

Correspondents of person locations (i.e., scores) versus item locations/difficulties (N = 3,413).

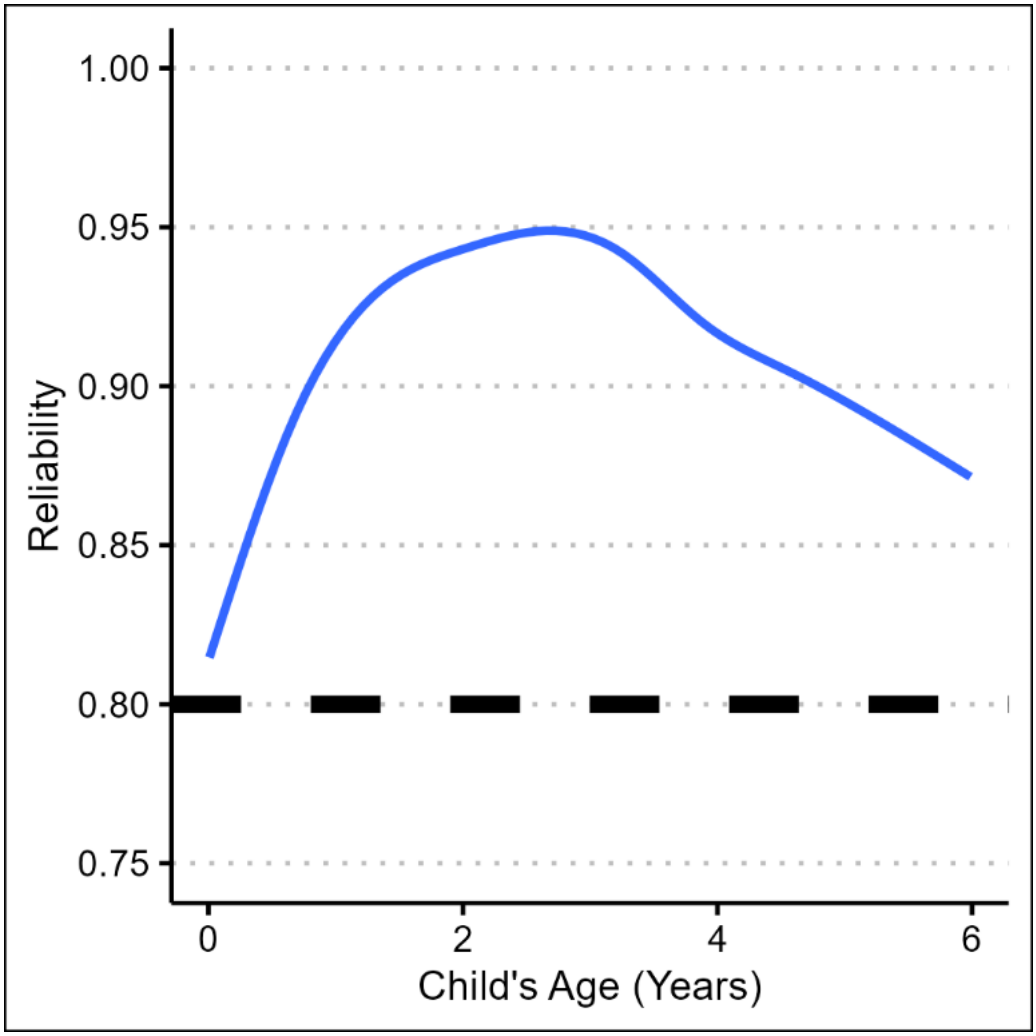

Figure 4

Average reliability of scores at child's age (N = 3,413).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 4supplementalmaterialsValidationKidsights.docx