



National Institutes of Health (NIH) Baby Toolbox®

Technical Manual

Y. Catherine Han, Courtney K. Blackwell, Elizabeth M. Dworak,
Rachel M. Flynn, Maxwell A. Mansolf, Miriam A. Novack, Sarah
Pila, Aaron J. Kaat

Northwestern University

Copyright

The NIH Baby Toolbox® is trademarked by the U.S. Department of Health and Human Services. The NIH Baby Toolbox® content and any derivatives of this work are copyrighted by Toolbox Assessments, Inc. All rights reserved. May not be reproduced in whole or in part in any form or by any means without permission.

Manual Version: NBT-TM1.1 Applicable to App Version 1

© 2025 Toolbox Assessments

Acknowledgements

We would like to acknowledge the efforts of the faculty and staff who have contributed countless hours to the development and analyses included in this NIH Baby Toolbox® Technical Manual (listed alphabetically by surname):

Hubert Adam, Greg J. Byrne, Richard C. Gershon, Cindy Nowinski, Abigail B. Sivan,
Vitali Ustsinovich, Lihua Yao

We would also like to acknowledge our software development and technology team at Northwestern University and Helium Foot for supporting the implementation of measures in the NIH Baby Toolbox® iPad platform (listed alphabetically by surname):

Christian Clare, Maria Varela Diaz, Cecily Frederick, Rachel Hanrahan, Pranav Rajiv,
Paula H. Sievert

We would also like to acknowledge the copyeditor of this technical manual, Anna Lane.

This project has been funded in whole or in part with federal funds from the National Institutes of Health Blueprint for Neuroscience Research and the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), and the National Institutes of Health, Department of Health and Human Services, under Contract No. 75N94019D00005.

Record of Revisions

Version	Date revised	Changes
1.0	2025-03-13	First release of the NIH Baby Toolbox technical manual
1.1	2025-06-18	Revision to description of Reach to Eat scoring (Chapter 2.3.2) and minor revision to description of statistics and time between test-retest (Chapter 4.1)

Table of Contents

Copyright.....	2
Acknowledgements	2
Record of Revisions	3
1. Overview of the NIH Baby Toolbox®	6
1.1 Background and History.....	6
1.2 NIH Baby Toolbox App Structure and Organization	7
1.3 Description of the Domains and Measures	7
1.3.1 Cognition Domain.....	7
1.3.2 Motor Domain.....	10
1.3.3 Social-Emotional Functioning Domain.....	12
1.4 Technical Manual Organization	14
2. Measure Development and Scoring.....	14
2.1 Scoring Algorithms	14
2.2 Cognition Domain.....	17
2.2.1 Familiarization.....	17
2.2.2 Visual Delayed Response (VDR).....	18
2.2.3 Delayed Memory.....	19
2.2.4 Mullen Visual Reception	21
2.2.5 MacArthur-Bates Communicative Development Inventories Computerized Adaptive Testing (CDI-CAT) Production	22
2.2.6 MacArthur-Bates Communicative Development Inventory Computerized Adaptive Testing (CDI-CAT) Comprehension	23
2.2.7 Looking While Listening (LWL)	24
2.2.8 Mullen Receptive	26
2.2.9 Mullen Expressive.....	27
2.2.10 Picture Vocabulary (PV)	28
2.2.11 Counting.....	28
2.2.12 Subitizing.....	29
2.2.13 Who Has More	30
2.2.14 Verbal Arithmetic.....	31
2.3 Motor Domain	31
2.3.1 Get Up and Go	31
2.3.2 Reach to Eat	35
2.4. Social-Emotional Functioning Domain	37
2.4.1 Caregiver Checklist	38
2.4.2 Social Observation Measure (SOM)	39
2.4.3 Infant, Early Childhood, and Childhood Behavior Questionnaires	40

2.4.4 Patient-Reported Outcomes Measurement Information System (PROMIS®) Early Childhood (EC) Parent Report Measures	42
2.5 Experimental/Supplemental Measures	44
2.5.1 Sit and Stand	44
3. Norming of the NIH Baby Toolbox	46
3.1 Description of the Norming Study	46
3.2 Norming Procedures.....	52
3.2.1 Norming of Individual Test and Composite Scores	52
4. Reliability and Validity of the NIH Baby Toolbox	54
4.1 Reliability	54
4.1.1 Empirical Reliability	54
4.1.2 Cognition Domain Reliability	54
4.1.3 Motor Domain Reliability	57
4.1.4 Social-Emotional Functioning Domain Reliability	58
4.1.5 Composites	61
4.2 Validity	63
4.2.1 Developmental Change in Measures and Composites by Age	63
4.2.2 Relationship of the NIH Baby Toolbox to other Measures	66
4.3 Factor Structure for Measure Scores	76
4.3.1 Measure and Composite Score Intercorrelations	76
4.3.2 Composite Development	82
4.3.3 Allowable Patterns and Composite Score Calculation	85
5. References.....	87

1. Overview of the NIH Baby Toolbox®

The NIH Baby Toolbox® (NBT) is a groundbreaking measurement tool designed for children aged 16 days through 42 months (Gershon et al., 2024). The NIH Baby Toolbox leverages the capabilities of an iPad, incorporating built-in gaze-tracking technology and brief video clips, to facilitate precise, objective assessments of children's development in domains including Cognition, Motor, and Social-Emotional Functioning. The NIH Baby Toolbox is designed to benefit researchers interested in investigating behavioral and neurological function in the youngest children and is also relevant in clinical and educational settings.

1.1 Background and History

To address the need for a standardized, nationally normed assessment suitable for infants and toddlers under 42 months old (3.5 years), the NIH Blueprint for Neuroscience Research, with management by the *Eunice Shriver Kennedy* National Institute of Child Health and Human Development (NICHD), solicited a contract in spring 2019 to develop an extension of the NIH Toolbox for Assessment of Neurological and Behavioral Function® (NIH Toolbox®; Gershon et al., 2013) for appropriate use for infants and toddlers. Northwestern University was awarded the contract and compiled a multi-institutional scientific team to produce a modern set of measures for developmental populations. The resulting developmental assessment, known as the NIH Infant and Toddler Toolbox, or “NIH Baby Toolbox,” is designed to provide cutting-edge assessment tools for children from 1-42 months old. It is envisioned as a “common currency” for research and clinical use—a comprehensive, nationally normed, and standardized battery of developmental functioning.

The goal of measure development for the NIH Baby Toolbox was to ensure the selected measures reflected the most up-to-date research in the field and identify and remedy any problems or additional developmental and validation work required for inclusion in the NIH Baby Toolbox. To accomplish this, the NIH Baby Toolbox research team utilized an expert survey of developmental researchers and clinicians, as well as a scoping literature review.

The NIH Baby Toolbox also leverages advanced approaches in item development, measure construction, and scoring. Much of the NIH Baby Toolbox is scored using item response theory (IRT), and several measures utilize computerized adaptive testing (CAT). Both methods allow measures to be brief, yet precise and valid. Using IRT, items from the NIH Baby Toolbox measures are calibrated along a continuum covering the full range of the construct being measured, which allows for the usage of CAT to identify the appropriate administration of items based on the individuals' responses. In doing so, burdens on the child, caregiver, and administrator are minimized while precise evaluation at the individual level still occurs.

The NIH Baby Toolbox is available in both English and Spanish. It includes brief measures across key developmental areas including memory, learning, executive functioning, language, math, social-emotional functioning, and motor. It was primarily intended to include direct assessments of infant and toddler behavior, accompanied by observation-based coding and caregiver report. The full battery is designed to be administered in person by researchers or clinicians in under 1 hour.

1.2 NIH Baby Toolbox App Structure and Organization

The NIH Baby Toolbox spans three domains or broad areas of developmental health and function: Cognition, Motor, and Social-Emotional Functioning. A measure is a set of items that are administered in an order determined by the app. The content of the items and the order of the item presentation cannot be changed within a measure. Each measure can be administered individually or as part of a preset battery.

The NIH Baby Toolbox batteries are intended to be administered in a specific order. However, individual measures can be added or removed from the assessment. The app produces scores for each measure in the battery as well as subdomain composites and domain-level composites for the Cognition, Motor, and Social-Emotional Functioning domains. However, please note that any missing measure-level scores towards subdomain or domain-level composites will result in missing composite scores.

Please refer to the *NIH Baby Toolbox Administration Manual* for detailed instructions on using the app.

1.3 Description of the Domains and Measures

1.3.1 Cognition Domain

Cognition during early development encompasses a wide range of subdomains, including Learning, Memory, Attention, Language, and even early number concepts (Math). These subdomains are foundational for children's development and are predictors of later outcomes like academic achievement (Blair & Razza, 2007; Hohm et al., 2007; Rose et al., 1991). Supplemental measures are not used in the calculation of composites but can be administered to better understand the participant's cognitive functioning.

1.3.1.1 Measures in the Cognition Domain

Measures within the Cognition domain can be found in Table 1.1, organized by subdomain.

Table 1.1*Description of NIH Baby Toolbox Cognition Domain Measures*

Subdomain	Measure	Age range (months)	Description	Modality	Time to administer (minutes)
Language	Looking While Listening	6-24	An assessment of object-labeling using a gaze-based preferential looking paradigm. Infants hear a prompt labeling one of two images presented on a screen.	Gaze	5
Language	Picture Vocabulary	≥25	An assessment of word knowledge. Children are asked to select one of four images via a touch-based measure.	Touch	3
Language	MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests (CAT) Production	9-30	An assessment of vocabulary production. Computerized Adapted Test version of the classic MCDI parent report and requires responses to only 20-50 words.	Caregiver Report	2
Language	MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests (CAT) Comprehension	7-18	An assessment of vocabulary comprehension. Computerized Adapted Test version of the classic MCDI parent report and requires responses to only 20-50 words.	Caregiver Report	1
Language	Mullen Receptive	1-42	An assessment of receptive language and communication skills based on elicited prompts.	Observation	3
Language	Mullen Expressive	1-42	An assessment of expressive language and communication skills based on elicited prompts.	Observation	3
Executive Function/ Memory	Mullen Visual Reception	1-42	An assessment of visual reception using both elicited prompts and touch-based responses on the iPad.	Observation and Touch (Age Dependent)	4

Executive Function/ Memory	Familiarization	6-21	An assessment of the ability to habituate to a visual stimulus.	Gaze	4
Executive Function/ Memory	Visual Delayed Response	22-42	An assessment of attention, executive functioning, and memory via a visual-delayed touch-based paradigm.	Touch	5
Executive Function/ Memory	Delayed Memory	22-42	A touch-based assessment of memory. Children are asked to encode items from the Memory Task (Learning) and provide recognition of individual items after a 6-8-minute delay.	Touch	4
Math	Who Has More	25-42	An assessment of the approximate number system by asking children to select which of two numerical quantities on a screen is larger.	Touch	3
Math	Subitizing	25-42	An assessment of the approximate number system and number recognition by asking children to report on the number of objects presented briefly on a screen.	Observation	1
Math	Verbal Arithmetic	37-42	An assessment of addition and subtraction by asking children to respond to simple arithmetic questions verbally.	Observation	2
Math	Counting	25-42	An assessment of counting and cardinality by asking children to count as high as possible and to report the number of objects presently on a screen.	Observation	2

1.3.1.2 Cognition Domain Batteries

Measure-level scores are available for all Cognition domain measures. Composite scores for Executive Function/Memory, Math, Motor, Mullen Language, and overall Cognition are available depending on successful completion of specific measures (see Table 1.2.). For user convenience, a preset battery of measures was created within the

app for the Cognition domain (see Table 1.3). Toddlers 25-42 months old who complete all measures within this battery can receive all possible composite scores listed in Table 1.2.

Table 1.2

Cognition Domain Composite Scores

Cognition domain composite scores available	Age range (months)	Measures included towards composite scores
Executive Function/Memory	22-42	Memory Task (Learning and Test), Visual Delayed Response
Math	25-42	Who Has More, Subitizing, Counting
Mullen Language	1-42	Mullen Expressive, Mullen Receptive
Cognition	25-42	Memory Task (Learning and Test), Visual Delayed Response, Who Has More, Subitizing, Counting, Mullen Expressive, Mullen Receptive

Table 1.3

Cognition Domain Preset Batteries Available to Users

Batteries	Age range (months)	Battery measures	Composite scores produced
Cognition (25-42 months)	25-42	Memory Task (Learning and Test), Visual Delayed Response, Who Has More, Subitizing, Counting, Mullen Expressive, Mullen Receptive	Executive Function/Memory Composite, Math Composite, Mullen Language Composite, Cognition Composite

1.3.2 Motor Domain

Motor function involves complex physiological processes and requires the integration of multiple systems, including neuromuscular, musculoskeletal, cardiopulmonary, neural motor, and sensory-perceptual systems. The NIH Baby Toolbox Motor domain measures the assessment of manual dexterity, manual planning, locomotion, postural transitions, and stationary postural control. The tests produce Fine Motor and Gross Motor scores as well as an overall Motor composite score. Supplemental measures are not used towards the calculation of composites but can be administered to better understand the child's motor functioning.

1.3.2.1 Measures in the Motor Domain

Measures within the Motor domain can be found in Table 1.4, organized by subdomain.

Table 1.4*Description of NIH Baby Toolbox Motor Domain Measures*

Subdomain	Measure	Age range (months)	Description	Modality	Time to administer (minutes)
Gross Motor	Get Up and Go	1-42	An assessment of postural transitions, locomotion, and overall gross motor ability by observing how children get up, travel 3 meters, and get onto and off a step platform.	Caregiver Report/Observation	4.5
Fine Motor	Reach to Eat	1-42	An assessment of fine motor dexterity, manual planning, and fine motor ability by observing how children grasp small objects and use tools.	Caregiver Report/Observation	5.5
Postural Control (Gross Motor, Supplementary)	Sit and Stand	1-42	An assessment of stationary postural control by observing children who are placed in various sitting and standing positions.	Caregiver Report/Observation	3.5

1.3.2.2 Motor Domain Batteries

The overall Motor composite scores (1-42 months old) require the completion of both the Get Up and Go and Reach to Eat measures. Specifically, the Fine Motor or Dexterity (Block) scores from the Reach to Eat measure and the Gross Motor score from the Get Up and Go measure are required towards computing the Motor composite score (see Table 1.5).

Table 1.5*Motor Domain Preset Batteries Available to Users*

Batteries	Age range (months)	Battery measures	Composite scores produced
Motor (1-42 months)	1-42	Get Up and Go, Reach to Eat	Motor Composite

1.3.3 Social-Emotional Functioning Domain

Social-Emotional Functioning refers to the social competencies related to the perception/recognition of others, social responsiveness/reciprocity, positive social interactions and communication, and positive relationships with adults and peers (Denham et al., 2009). It also includes the experience, expression, and management of any strong feelings (i.e., emotions), such as excitement, fear, or anger, which can be positive or negative (Denham et al., 2009). Early social-emotional functioning is foundational for children's health and development, and a meaningful predictor of school readiness, academic achievement, social competence, and psychopathology (Baumeister & Leary, 1995; Denham, 2006; Duncan et al., 2007; Eisenberg et al., 2000, 2001; Ghandour et al., 2019; Jones et al., 2015; McClelland et al., 2007; Moffitt et al., 2011; Osher et al., 2021; Rose-Krasnor, 1997; Rothbart et al., 1994).

1.3.3.1 Measures in the Social-Emotional Functioning Domain

Measures within the Social-Emotional Functioning domain can be found in Table 1.6, organized by subdomain.

Table 1.6

Description of NIH Baby Toolbox Social-Emotional Functioning Domain Measures

Subdomain	Measure	Age range (months)	Description	Modality	Time to administer (minutes)
Temperament	Infant Behavior Questionnaire - Revised - Very Short Form (IBQ-R-VSF)	3-12	An assessment of infant's effortful control, surgency, negative affectivity, and fear of unfamiliar adults.	Caregiver Report	6
Temperament	Early Childhood Behavior Questionnaire - Very Short Form (ECBQ-VSF)	13-36	An assessment of young children's effortful control, surgency, and negative affectivity.	Caregiver Report	6
Temperament	Children's Behavior Questionnaire - Very Short Form (CBQ-VSF)	37-42	An assessment of children's effortful control, surgency, and negative affectivity.	Caregiver Report	6

Negative Affect	PROMIS EC Parent-Report Bank v1.0 - Anger/Irritability	13-42	An assessment of children's angry mood, including irritability and grouchiness, and angry behavior, such as tantrums.	Caregiver Report/CAT	1
Negative Affect	PROMIS EC Parent-Report Bank v1.0 - Anxiety	13-42	An assessment of children's feelings of fear, anxious misery, hyperarousal, and social/separation anxiety.	Caregiver Report/CAT	1
Negative Affect	PROMIS EC Parent-Report Bank v1.0 - Depressive Symptoms	13-42	An assessment of children's feelings of sadness and withdrawal, negative self-views, and anhedonia.	Caregiver Report/CAT	1
Psychological Well-Being	PROMIS EC Parent-Report Bank v1.0 - Positive Affect	13-42	An assessment of children's mood and feelings associated with positive affective experiences, such as happiness, joy, and enthusiasm.	Caregiver Report/CAT	1
Self-Regulation	PROMIS EC Parent-Report Scale v1.0 - Self-Reg - Flexibility 5a	13-42	An assessment of children's ability to adapt in response to environmental demands, changes, and expectations.	Caregiver Report	1
Self-Regulation	PROMIS EC Parent-Report Scale v1.0 - Self-Reg - Frustration Tolerance 6a	13-42	An assessment of children's ability to recognize and regulate emotions and behaviors to serve one's own goals.	Caregiver Report	1
Social Relationships	PROMIS EC Parent-Report SF v1.0 - Soc Rel - Child-Caregiver Interactions	13-42	An assessment of positive child-caregiver interactions, experiences, and connectedness.	Caregiver Report	1
Social Relationships	PROMIS EC Parent-Report SF v1.0 - Soc Rel - Peer Relationships 4a	13-42	An assessment of children's positive interactions with peers, sociability, and empathic behaviors.	Caregiver Report	1
Social Relationships, Social Communication	Caregiver Checklist	6-42	An assessment of children's interactions with people and the environment.	Caregiver Report	2
Social Relationships, Social Communication	Social Observation	9-42	An assessment of children's social functioning, including social attention, social communication, play, and prosocial behavior, in response to specific, standardized verbal prompts and probes, and during more spontaneous social interactions with the caregiver.	Observation	10

1.3.3.2 Social-Emotional Functioning Domain Batteries

Measure-level scores are available for all measures in Table 1.6. Composite scores are available for Negative Affect, Self-Regulation, and Social Communication depending on the successful completion of specific measures, (see Table 1.7). No preset battery is available for the Social-Emotional Functioning domain, but individual measures can be selected towards receiving the composite scores in Table 1.7.

Table 1.7

Social-Emotional Functioning Domain Composite Scores

Social-emotional domain composites	Age range (months)	Measures towards composite
Negative Affect	13-42	Early Childhood Behavior Questionnaire - Very Short Form (ECBQ-VSF) or Children's Behavior Questionnaire - Very Short Form (CBQ-VSF), PROMIS EC Parent-Report Bank v1.0 - Anger/Irritability, PROMIS EC Parent-Report Bank v1.0 - Anxiety, PROMIS EC Parent-Report Bank v1.0 - Depressive Symptoms
Self-Regulation	13-42	Early Childhood Behavior Questionnaire - Very Short Form (ECBQ-VSF) or Children's Behavior Questionnaire - Very Short Form (CBQ-VSF), PROMIS EC Parent-Report Scale v1.0 - Self-Reg - Flexibility 5a, PROMIS EC Parent-Report Scale v1.0 - Self-Reg - Frustration Tolerance 6a
Social Communication	9-23	Caregiver Checklist, Social Observation

1.4 Technical Manual Organization

The following chapters include important technical information about the NIH Baby Toolbox measures and batteries. Chapter 2 focuses on the development of the NIH Baby Toolbox measures and detailed information about the scoring of the measures within each domain. Chapter 3 describes the norming studies conducted for the NIH Baby Toolbox domains. Chapter 4 contains information regarding the reliability and validity of the measures and batteries in the NIH Baby Toolbox. References are included in Chapter 6.

2. Measure Development and Scoring

2.1 Scoring Algorithms

In the NIH Baby Toolbox, the dichotomous and graded response model (GRM) engines are leveraged that exist within the NIH Toolbox to support scoring measures with

dichotomous or polytomous items. The dichotomous engine uses common IRT methods, including maximum information item selection and maximum likelihood scoring methods (Baker & Kim, 2004). The GRM engine was originally designed as part of the Patient Reported Outcomes Measurement Information System (PROMIS®) and implemented in the HealthMeasures Assessment Center (Gershon et al., 2010). It uses maximum posterior weighted information item selection and expected a posteriori (EAP) scoring.

In addition, a new multidimensional IRT calculator has been introduced, called the maximum a posteriori (MAP) calculator after the eponymous IRT scoring algorithm. While the dichotomous and GRM engines use quadrature nodes to integrate over the measured latent variable, this is not computationally feasible when multiple latent variables (dimensions in latent space) are measured, as in, for example, Social-Emotional Functioning measures including IBQ-R-VSF and SOM–Younger. This is because the number of quadrature nodes required scales exponentially with the number of dimensions. To demonstrate, the unidimensional GRM has 81 quadrature nodes, but with two dimensions $81^2 = 6,561$ nodes would be needed, $81^3 = 531,441$ for three, etc. To avoid this computational burden, the MAP calculator uses the Newton-Raphson Method (an iterative search procedure), to search the latent space for the combination of latent trait values (one per dimension) that maximize the posterior likelihood of obtaining the observed item response pattern.

Specification for the MAP calculator differs from dichotomous and GRM engines in several specific ways. As noted above, the MAP calculator does not require specifying the number or location of quadrature nodes, because it does not use them. Instead, measure-specific boundaries are placed on the search to prevent implausibly large or small scores from being generated. Unlike the GRM or dichotomous engines, the MAP calculator can accommodate combinations of dichotomous (with or without incorporating guessing) and graded (i.e., polytomous) items within the same test.

Just as the GRM engine incorporates a prior distribution in applying EAP scoring, the MAP calculator incorporates a prior distribution, but in the MAP calculator, this is a multivariate distribution to account for multiple latent variables measured. One feature of this prior is it can include correlations between latent variables, with two consequences. First, when different item sets measure different latent variables, and each set has data to use for scoring, measurement precision (i.e., reliability) on both latent traits is increased. Second, scores on each latent variable can differ slightly depending on responses to items not directly measuring that latent variable. Both properties arise from the assumption, operationalized by the prior, that scores on each latent dimension are related to some degree. Note that this last property does not apply when unidimensional measures use the MAP calculator due to its ability to accommodate mixed item types (e.g., Counting) or when the multidimensional model has uncorrelated latent variables

(e.g., Caregiver Checklists); these measures still incorporate a prior, but all item responses inform on the single reported score in both cases.

2.2 Cognition Domain

This chapter describes the goals and objectives, development, and scoring approaches for the measures in the Cognition domain of the NIH Baby Toolbox, which encompasses Executive Function and Memory (EF-Mem), Language, and Math/Numeracy subdomains.

Executive Function and Memory

2.2.1 Familiarization

The Familiarization measure is a gaze-based measure of infant habituation to a repeatedly presented visual stimulus (Colombo & Mitchell, 2009), which assesses learning and speed of processing for children under 24 months old. Visual habituation has been studied for a long time, with the first seminal study often attributed to Fantz (1964). It has been shown in literature that Familiarization has significant predictive power for later cognitive development, such as how quickly the child habituates to a stimulus as being predictive of their future cognitive abilities (Colombo et al., 2004; McCall & Carriger, 1993).

Measure Development

The Familiarization measure was added to the NIH Baby Toolbox as a new gaze-based measure for children too young or not capable of touch-based responses on an iPad. The simplest habituation protocols have been shown to use fixed trial procedures, which involve administering a set of discrete, repetitive stimulus presentations to an infant with a fixed trial duration and fixed intertrial interval. A common protocol includes eight trials of a checkerboard to an infant, each lasting 10 seconds (e.g., Colombo, 1997), which was utilized in the implementation of this measure.

Administration and Scoring

The Familiarization measure progresses through two different rounds. In the first round, a dog is shown in the center of the iPad screen for 10 seconds while music plays. This is repeated for eight total items to familiarize the child to the animal. The second round proceeds the same way as the first round with a pig as the familiar item, which is also repeated for eight items. Item-level scoring was derived as the duration of looking (in seconds) to the center of the screen where the animal is located.

Several types of measure-level scores are available, which include peak looking duration, summed looking duration, and mean looking duration in seconds for Round 1, Round 2, as well as across Both Rounds. If the child looked away from the center of the screen entirely during a round (i.e., 0 seconds of looking to the center for all eight items within a round), the child will not receive a Both Rounds score.

2.2.2 Visual Delayed Response (VDR)

The Visual Delayed Response (VDR) measure is a touch measure of EF, memory, and speed of processing (Pelphrey et al., 2004) for children aged 22-42 months. This measure involves hiding objects in one or more locations, followed by distracting the child to break their attention. An unnameable creature appears in the middle of the screen and then hides between either a left or right box. The curtain closes and music plays, the duration for which is age-based increasing in difficulty. The child is instructed to touch the box for which the creature hid, and the creature appears when the correct box is touched.

Measure Development

The VDR measure was added to the NIH Baby Toolbox to address the need for a measure with a higher order of EF skills for young children who are capable of a touch-based response on an iPad.

Administration and Scoring

Prior to completion of any touch measure, including VDR, the participant should have completed the NIH Baby Toolbox Touchscreen Tutorial. The participant completes three practice items, for which the box moves from transparent to partially (but not completely) opaque. In this phase, the app provides feedback for correct and incorrect responses. Next, the participant completes eight live items, the latter four of which have a longer curtain delay time to increase measure difficulty. The app captures the participant's reaction time to touch down, item response, and item score.

The raw item-level score for VDR is determined based on whether the examinee answered correctly for the item. Items were calibrated to a 2-parameter logistic (2PL) multiple group IRT model, where item discrimination and difficulty varied in relation to the underlying EF trait. Examinees were scored using maximum likelihood scoring to obtain an interval-scaled score on the logit metric, which is then converted to a change-sensitive score (CSS) for reporting purposes.

Additional scores produced by this measure include scores based on reaction time, including the Summed Residual Time (Equation 2.1; Maris & van der Maas, 2012; van der Maas & Wagenmakers, 2005) and Median Reaction Time, which are available across (1) correct trials, (2) incorrect trials, (3) and all correct and incorrect trials. Additionally, for (2) and (3), a version of these scores is available that accounts for missed responses, such that these scores are computed with only correct or incorrect touch responses recorded within the response time limit.

Equation 2.1: $\sum_{i=1}^n (2X_i - 1)(d - T_i)$,

where i is defined as the current trial, X_i is the trial-level accuracy, d is the trial time limit, and T_i is the trial-level response time.

2.2.3 Delayed Memory

The Delayed Memory measure evaluates learning and memory in infants 22-42 months old via touch-based responses. This measure is designed to ensure that memory is not confounded with inhibitory control. This is accomplished by having the same inhibition demands during both the encoding and test, and by having corrective feedback during the encoding portion of the measure. Providing corrective feedback ensures young children will learn the rule to touch the novel animal, ensuring that accuracy during the test phase is due to memory.

Measure Development

The Delayed Memory measure uses appealing and familiar animals as the stimuli and a narrative game aspect of feeding hungry animals to make it meaningful to children. Prior research demonstrates that children perform better with narrative pretense rather than abstract concepts because it reduces the cognitive load (Carlson et al., 2014). Animals were selected using Wordbank's database (Frank et al., 2017; Kachergis et al., 2022) based on the percentage of English-learning children who produce the word (animal) at 24 months old since word comprehension would likely be high among children who can produce the word.

Administration and Scoring

Prior to completion of any touch measure, including Delayed Memory, the participant should have completed the NIH Baby Toolbox Touchscreen Tutorial. Prior to the learning portion, infants are given three practice trials in which animal images are introduced. Infants are instructed to "touch the animal to feed it" on each practice trial and are given corrective feedback. Practice trials are repeated if there are errors. During the learning portion, infants are told: "Now you see two animals. One animal we already fed. Let's touch the NEW animal." They are then exposed to 11 animal images in which the animal seen on the previous trial is presented with a novel animal on the left and right sides of the screen. For each trial, the infant is again instructed to "touch the NEW animal" by touching the novel animal image of the image pairs for learning purposes. They receive corrective feedback on each trial: "Yum! That's right!" or a wrong answer sound paired with: "We fed that one already, remember! Touch the new animal." During the memory portion of the measure, which is after a delay of 6-8 minutes, the examinee is presented with 20 images that formed 10 dyads, where the animals from the learning portion are paired with two different novel animals. This ensures that the animals they learned are seen twice with two different novel animals to reduce preference towards

certain animals. After each item in the test phase, children receive only neutral feedback with: “Let’s do another.”

Memory Learning

The practice trials are scored as correct or incorrect based on the examinee’s response. The 11 items were calibrated to a 2PL IRT model, where item discrimination and difficulty varied in relation to the underlying trait. Examinees were scored using maximum likelihood scoring to obtain an interval-scaled score on the logit metric, which is then converted to a CSS for reporting purposes.

Additional scores produced by this measure include scores based on reaction time. These include the Summed Residual Time (see Equation 2.1) and Median Reaction Time, which are available across (1) correct trials, (2) incorrect trials, (3) and all correct and incorrect trials. Additionally, for (2) and (3), a version of these scores is available that accounts for missed responses, such that these scores are computed with only correct or incorrect touch responses recorded within the response time limit.

Memory Test

The Memory Test is administered 6-8 minutes after the learning portion of the measure. During the delay, VDR is typically administered. Items consisted of dyads, in that images presented during the first 10 items of the test (and thus presented during the learning portion) were also presented again within the second 10 items of the test. For example, an image of a dog was presented during the learning portion and was presented once during the first half as part of an image pair (dog and cat) and again during the second half as part of a different image pair (duck and dog).

Dyad items were scored as correct if both items in the dyad were correct or as incorrect if only one or none of the dyad items were correctly chosen. This was implemented to provide greater assurance that the child remembered the item rather than random responding. The 10 dyad items were calibrated to a 2PL IRT model, where item discrimination and difficulty varied in relation to the underlying delayed memory trait. Examinees were scored using maximum likelihood scoring to obtain an interval-scaled score on the logit metric, which was then converted to a CSS for reporting purposes.

Additional scores produced by this measure include scores based on reaction time. These include the Summed Residual Time (see Equation 2.1) and Median Reaction Time, which are available across (1) correct trials, (2) incorrect trials, (3) and all correct and incorrect trials. Additionally, for (2) and (3), a version of these scores is available that accounts for missed responses, such that these scores are computed with only correct or incorrect touch responses recorded within the response time limit.

2.2.4 Mullen Visual Reception

The Mullen Visual Reception (MVR) measure is designed to measure a child's performance in processing visual patterns (the input side of visual organization skills). The MVR measure is intrasensory; it presents visual information in various forms and patterns and assesses visual processing skills. Because the measure is intrasensory, objects and pictures are not labeled. The items adapted for the NIH Baby Toolbox involve oculomotor and visual-motor operations (e.g., localizing a target, visual tracking, or scanning multiple points on a surface). The Mullen Early Scales of Learning (Mullen) was originally developed based on expert conceptualizations of development and assessment, including a priority on minimizing language requirements to understand and complete the MVR measure. MVR measures early cognitive skills by emphasizing visual perceptual ability (input decoding) while minimizing output—or response—requirements. Motor response requirements are minimal, and verbal responses are not required. Except for the earliest items, the child responds by the simple manipulation of objects (e.g., matching) or by pointing.

Measure Development

The MVR measure is made up of a subset of the items adapted from the Mullen Scales of Early Learning Visual Reception Scale (Mullen, 1995) with permission from NCS Pearson, Inc. These items were chosen because they were amendable to iPad-based scoring and administration. The resulting NIH Baby Toolbox Mullen Visual Reception item bank includes 17 items.

Administration and Scoring

The starting item is administered based on the child's age. An incorrect item is defined as an item in which the child fails to score a single point. A correct item is defined as an item in which the child successfully scores at least one point (Please note: some items are worth multiple points; a child only needs to obtain partial credit on these items to establish a basal score). For items with multiple response screens (e.g., MVR27), all screens must be answered to obtain a score and advance to the next item. Some of these items require all screens to be correct, while others may provide partial credit and multiple points per item. If a child responds incorrectly to the first item, they are presented with the next easiest item until they respond correctly to an item or the first item on the measure is administered (i.e., the basal rule). If a child answers the first item that they are presented with correctly, a basal rule is immediately established, and they progress to the next item. They will continue to progress through the instrument until they fail to answer two consecutive items correctly or reach the last item in the measure (i.e., the ceiling rule). The iPad ensures item presentation is appropriate and will select the easier or harder item until the basal and ceiling rules are met.

Note that, although MVR is one instrument, it has two sections relevant to the administrator: The first section primarily includes items where the examiner and examinee are interacting together, and the examiner is scoring the behaviors; and the second section primarily includes items where the examinee is interacting with and responding to items presented on the iPad. As such, if a younger examinee begins with the first section and has not met a ceiling rule before entering the second section, the examiner may want to delay the measure, administer the Touchscreen Tutorial, and resume the MVR measure after the delay. This is not relevant if the Touchscreen Tutorial has already been administered.

Finally, to score the MVR measure, it is necessary for the child to have met a basal and a ceiling condition—that is, that they got at least one item correct (or took the easiest item on the measure) and had two consecutive items incorrect (or took the hardest item on the measure). The polytomous items were calibrated to a graded response IRT model, while dichotomous items were calibrated using a 2PL model. Given that there are mixed item types within the MVR measure, scores are calculated using the MAP calculator (Section 2.1), based on the items administered and scored between the basal and ceiling rules.

Language

2.2.5 MacArthur-Bates Communicative Development Inventories Computerized Adaptive Testing (CDI-CAT) Production

The MacArthur-Bates Communicative Development Inventories Computerized Adaptive Testing (CDI-CAT) Production measure assesses infants' productive vocabulary through caregiver reports. This inventory is designed for use with children aged 9-30 months and was adapted for the NIH Baby Toolbox with permission from the MacArthur-Bates CDI Advisory Board.

Measure Development

The original MacArthur-Bates Communicative Development Inventories (Fenson et al., 1994) are some of the most widely used caregiver reports for assessing vocabulary knowledge in young children (Frank et al., 2021).

A CAT version of the MacArthur-Bates CDI Production measure has been included in the NIH Baby Toolbox to reduce the testing burden compared to the original inventory. The original MB-CDI Production inventory contains 679 items for both English-speaking and Spanish-speaking children. The item parameters and CAT engine settings are consistent with those originally provided by the MacArthur-Bates CDI Advisory Board and implemented in the Web-CDI CAT, which is managed by the Advisory Board and distributed by other providers.

Administration and Scoring

Caregivers are administered items one at a time. For each item, caregivers are asked to indicate whether their child says the word.

The live test items are administered via a CAT algorithm. An appropriate starting item is selected based on the child's age in months. For both English and Spanish versions, the CAT administers a minimum of 20 items and a maximum of 50 items with a target probability of correct response of 0.675. Once the CAT has reached the minimum test length, the test will end if the standard error of the ability estimate falls below 0.15 logits.

Scoring for the English and Spanish measures follow a 2PL IRT model, with item parameters provided by the MacArthur-Bates CDI Advisory Board to ensure consistency across authorized distribution of the CDI-CAT. Item parameters were originally derived from data within the CDI Wordbank (Frank et al., 2017).

2.2.6 MacArthur-Bates Communicative Development Inventory Computerized Adaptive Testing (CDI-CAT) Comprehension

The MacArthur-Bates Communicative Development Inventory Computerized Adaptive Testing (CDI-CAT) Comprehension measure assesses infants' receptive vocabulary based on caregiver reports. This inventory is designed for children aged 7-18 months and was adapted for the NIH Baby Toolbox with permission from the MacArthur-Bates CDI Advisory Board.

Measure Development

The original MacArthur-Bates Communicative Development Inventories (Fenson et al., 1994) have long been one of the most widely used tools for measuring vocabulary knowledge in young children through caregiver reports (Frank et al., 2021).

The original MacArthur-Bates CDI Comprehension inventory for English-speaking children contains 396 items, and the inventory for Spanish-speaking children contains 428 items. The item parameters and the CAT engine settings are consistent with those originally provided by the MacArthur-Bates CDI Advisory Board and implemented in the Web-CDI CAT, which is managed by the Advisory Board and distributed by other providers.

Administration and Scoring

Caregivers are administered items one at a time and are asked to indicate whether their child understands each word.

The live test items are administered via a CAT algorithm. An appropriate starting item is selected based on the child's age in months. For both English and Spanish measures, the CAT administers a minimum of 20 items and a maximum of 50 items with a target probability of a correct response of 0.675. Once the CAT has reached the minimum test length, the test will end if the standard error of the ability estimate falls below 0.224 logits (corresponding to a person-level reliability estimate of approximately 0.95).

Scoring for the English and Spanish measures follows a 2PL IRT model, with item parameters provided by the MacArthur-Bates CDI Advisory Board to ensure consistency across authorized distribution of the CDI-CAT. Item parameters were originally derived from data within the CDI WordBank (Frank et al., 2017).

2.2.7 Looking While Listening (LwL)

The Looking While Listening (LwL) measure assesses comprehension for object labeling, or receptive language. Infants are presented with two images (e.g., a car and a shoe) while they hear a spoken sentence that labels one of the images (e.g., “Look at the car!”). The infant's gaze is tracked in real time to determine whether they look at the labeled referent.

Measure Development

The LwL measure was developed based on the widely used and robust LwL paradigm, designed to assess language comprehension and processing in infancy (e.g., Fernald et al., 2008). Drawing on the visual world paradigm (Tanenhaus et al., 1995), LwL takes advantage of infants' natural tendency to focus their gaze on a visual referent (e.g., a car) upon hearing its label in speech (e.g., “Look at the car!”). This measure has been employed with infants as young as 6 months old, (Bergelson & Swingley, 2012) and by 18 months old, it provides measures that reliably predict later language and cognitive development (Marchman & Fernald, 2008).

Target words were selected from the Wordbank database, which is based on the MacArthur-Bates Communicative Development Inventory. The words were chosen for their ease of recognition by young children (e.g., “car” or “shoe”) and for being highly imageable nouns (e.g., “car” but not “no”). Words were selected and matched based on age of acquisition and difficulty level. Additionally, the words were paired to ensure distinct starting sounds and different semantic categories. The list was reviewed by domain experts and adjusted for cultural relevance. A parallel set of target Spanish words was selected using the same criteria, with particular attention given to gender matching and consistency across Spanish-speaking regions. Seven Spanish items were also included in the English version.

For the audio stimuli, four unique carrier phrases were recorded for each word by a female voice actress using infant-directed speech. The Spanish recordings were made

by a different female voice actress, chosen by a group of native Spanish speakers for her clear, neutral tone that would be easily understood by speakers from various Spanish-speaking countries.

The visual stimuli were created by a professional illustrator, with input from domain scientists. The images were carefully refined to ensure clarity, cultural sensitivity, and visual distinctiveness. For the Spanish version, an additional review by a Spanish-language expert ensured that the images were culturally and linguistically appropriate. Pilot testing with children was conducted to assess the clarity and appeal of the images, with adjustments made as needed to optimize their engagement and effectiveness.

Administration and Scoring

In total, there are 40 items divided into five groups of a priori determined difficulty (eight items in each group). Of these 40 items, only 24 items are administered in a fixed sequence to each participant. The trials are divided by age group: younger children (6-14 months old) and older children (15-24 months old). Participants younger than 15 months old are given items from groups 1-3, while those 15 months old or older receive items from groups 3-5. Each image appears twice, paired with different items, but the image is only the target once per measure. The target images alternate between the left and right sides of the screen in a semirandom order, ensuring that they do not appear in consecutive trials.

Each trial starts with a 2-second baseline period during which the images are shown without audio to gauge the child's initial preference. Following this, an audio prompt directs the child to look at the target image. Gaze analysis begins once the audio finishes, ensuring that scores reflect the child's gaze response after hearing the target word. Two areas of interest (AOIs) are defined around the target and distractor images. Gaze is recorded at 60 frames per second, and the data is analyzed to determine the percentage of time spent looking at the target relative to the total time the images are on-screen. A trial is considered "correct" if the child spends more than 50% of the on-screen time looking at the target image, and "incorrect" if 50% or less of their gaze is directed at the target or the wrong stimuli. Trials in which the child spends more than 50% of the 6-second analysis window off-screen are excluded from scoring.

Scoring for the English LwL measure follows a 2PL IRT model, with item parameters based on two-thirds of the norming data. A subset of the English items is considered experimental (10 total). While they are administered to children and included in item-level exports, they are not part of the instrument scoring. An eligibility rule is applied to determine whether IRT scores can be produced. Specifically, at least six non-experimental items must have an item score value—whether all correct, all incorrect, or mixed cases—to calculate IRT values.

The Spanish LwL measure uses the same trial-level scoring criteria (e.g., the trial is considered “correct” if the child spends more than 50% of the on-screen time looking at the target image, and “incorrect” if 50% or less of their gaze is directed at the target or the wrong stimuli, and trials in which the child spends more than 50% of the 6-second analysis window off-screen are excluded from scoring). However, in contrast to the English version, the Spanish version uses a Percentage Correct score, which calculates the percentage of correctly completed items out of all completed items. The difference in scoring methodology is due to limited data collected during norming to create a valid IRT-based score for the Spanish version of this measure. A minimum of six items, whether all correct, all incorrect, or mixed, must have an item score value in order to calculate the Percentage Correct values.

2.2.8 Mullen Receptive

The Mullen Receptive language measure assesses receptive language through responses to prompted interactions (e.g., “Give the [toy] to [mommy/daddy]”). The Mullen Scales of Early Learning has a long history of use in developmental assessment (Mullen, 1995), and items were adapted for the NIH Baby Toolbox with permission from NCS Pearson, Inc. The NIH Baby Toolbox Mullen Receptive language measure was selected because it is relatively simple to administer and relies on observational responses, which aligns with one of the key goals of the NIH Baby Toolbox.

Measure Development

Items from the original Mullen Receptive language subscale were excluded from the NIH Baby Toolbox version if they 1) were determined to test skills other than language, 2) were intended for children beyond the NIH Baby Toolbox age range, or 3) if their content was not amendable to iPad-based scoring or administration. The resulting NIH Baby Toolbox Mullen Receptive language item bank includes 13 adapted items. Pilot testing of the original 13 items developed for the NIH Baby Toolbox indicated that one item was too difficult to administer. It was subsequently dropped from scoring, resulting in a final item bank of 12 items.

Administration and Scoring

Participants are administered a limited number of items with varying starting items based on the child’s age and adaptive administration that advances if the child answers a question correctly and reverts to a preceding unanswered item if the child answers a question incorrectly. The measure ends early if the child answers two prompted items incorrect in a row or when they reach the most difficult question.

Scoring for the English and Spanish measures follow a mixed IRT model using both a 2PL model and a graded response model. Given the mixed item types, the MAP calculator (Section 2.1) is used to derive IRT-based scores.

2.2.9 Mullen Expressive

The NIH Baby Toolbox Mullen Expressive language measure assesses expressive language via a combination of responses to prompted interactions (e.g., getting the child to label a ball) and observed behaviors (e.g., the examiner indicates whether they observed the child produce a two-word utterance at any point during the full NIH Baby Toolbox battery). The Mullen Scales of Early Learning has a long history of use in developmental assessment (Mullen, 1995), and items were adapted for the NIH Baby Toolbox with permission from NCS Pearson, Inc. The NIH Baby Toolbox Mullen Expressive language measure was selected because it is relatively simple to administer and relies on observational responses, which was one of the goals of the NIH Baby Toolbox.

Measure Development

The Mullen Expressive language measure is divided into two portions: “Prompted” and “Observational” to ease administration burden. All prompted items are administered together. All observed items are entered at the end of the full NIH Baby Toolbox battery. Some items from the original Mullen Expressive language subscale were excluded from the NIH Baby Toolbox version if they 1) were determined to test skills other than language, 2) were intended for children beyond the NIH Baby Toolbox age range, or 3) if their content was not amendable to iPad-based scoring or administration. The resulting NIH Baby Toolbox Mullen Expressive language measure item bank includes 17 adapted items. This includes five items in the prompted portion and 12 items in the observational portion.

Administration and Scoring

For the prompted portion of the Mullen Expressive language scale, participants are administered a limited number of items with varying starting items based on the child’s age. Adaptive administration advances if the child answers a question correctly and reverts to a preceding unanswered item if the child answers a question incorrectly. The measure ends early if the child answers two prompted items incorrect in a row or when they reach the most difficult question.

For the observational portion, examiners indicate which expressive language items were observed at some point during the testing session. Items later on the list are inclusive of earlier items.

Scoring for the English and Spanish versions of the measure follow a mixed IRT model using both a 2PL model and a graded response model. To get a score for the Mullen Expressive language measure, both the “Prompted” and “Observational” portion of the measure must be completed. Given the mixed item types, the MAP calculator (Section 2.1) is used to derive IRT-based scores.

2.2.10 Picture Vocabulary (PV)

The Picture Vocabulary (PV) measure assesses receptive language vocabulary and is identical to the NIHTB Picture Vocabulary Test, which was originally developed as part of the NIHTB Cognition battery (Gershon et al., 2014; Weintraub et al., 2013).

Measure Development

The NIHTB PV measure was already developed and validated in the general population for children as young as 3 years old. To extend the measure into the NIH Baby Toolbox age range, PV was pilot-tested with a sample of 2-year-old children prior to the NIHTB V3 release. Results confirmed that the measure was feasible for children at this age—that is, they were capable of completing the measure without exhibiting floor effects with the shared stimuli.

Administration and Scoring

Prior to completion of any touch measure, including PV, the participant should have completed the NIH Baby Toolbox Touchscreen Tutorial. PV uses the same dichotomous CAT engine for administration and scoring as the NIHTB V3. Participants are shown four images on the screen and asked via prerecorded audio prompts to touch the named object. The measure is administered as a CAT, with starting items based on participant age. Participants must complete at least 20 items, and the measure concludes after they have completed a maximum of 35 items or if they reach an SE of 0.44.

Scoring for the English and Spanish PV measures follow a 2PL IRT model, with item parameters provided by the NIHTB V3.

Math and Numeracy

2.2.11 Counting

The Counting measure is administered in two parts to children aged 25 months or older: Verbal Counting and Object Counting. Verbal Counting was designed to assess a child's ability to verbalize their knowledge of numbers in order. Object Counting was designed to assess a child's ability to look at an array of objects on a screen and count them.

Measure Development

Counting and cardinality is a foundational math skill for young children. Counting in order allows children to practice the concept of numbers, sequence, and quantity while cardinality—the understanding that the last number counted refers to the total number of items—facilitates number recognition and quantity comparison (another construct assessed by the math measures). Together, these skills not only contribute to future

math proficiency, but also cognitive abilities at large. Above all else, one goal of the NIH Baby Toolbox was to assess early math skills (and other respective domains of development) in under 10 minutes. As such, the measure is expected to capture all counting and cardinality skills in a few short minutes of testing.

Administration and Scoring

For the Verbal Counting part of the measure, the examiner asks the participant to count as high as they can starting from 1 and records the highest digit number that was serially provided. For the Object Counting part of the measure, the examiner presents a series of object arrays to the child on the iPad screen and enters the child's numeric response. For this part, there are six stimulus arrays, and the measure is automatically terminated if the child provides incorrect/no response(s) on the first (one camera) or third item (two trees). Otherwise, they receive a score for all items.

Items from the Verbal Counting part of the measure were scored categorically, such that as the highest counted value of 0 was scored as 0, the highest count of 1 or 2 was scored as 1, the highest count of 3, 4, or 5 was scored as 2, and any counts higher than 5 were scored as 3. Items from the Object Counting part of the measure were dichotomous (0 for incorrect responses, 1 for correct responses). The Verbal Counting item was calibrated using a GRM, whereas the Object Counting items were calibrated using a 2PL model. Given the mixed item types, the MAP calculator (Section 2.1) is used to derive IRT-based scores.

2.2.12 Subitizing

The Subitizing measure is administered to children aged 25 months or older and assesses a child's ability to detect the exact number of items on a screen shown for a short period of time.

Measure Development

Subitizing, the rapid and accurate recognition of the number of items in a small set without counting, is crucial for measurement even in the earliest years. This skill enables individuals to instantly assess quantities, which is fundamental in various measures that require quick judgments, such as choosing the most filled container or identifying the larger group of objects at a glance. It is also essential for cognition more generally. Given the inherently quick nature of capturing this skill, subitizing was an excellent candidate for measurement with just a few items in the allotted time.

Administration and Scoring

A dot image display is presented on the iPad screen for 1 second, and the child is prompted to say how many they see. The administrator records their response. There are four items of small quantities presented (e.g., 1-4) in random order.

Items were dichotomous, in which a 0 is given for incorrect responses, and a 1 is given for correct responses. Exploratory correlational analyses were conducted to determine the dimensionality of the items in relation to each other. The four items were used to construct confirmatory IRT models, and the data was fitted to a unidimensional Rasch (1PL) response model and scored using maximum likelihood (ML) score estimation.

2.2.13 Who Has More

The Who Has More measure is administered to children aged 25 months or older and assesses a child's ability to make quick estimations without time to count. Experts identified the innate ability to estimate and compare quantities without relying on language or symbols as an important one to evaluate. Measuring this capability—approximate number sense (ANS)—provides valuable insights into how individuals perceive and process numerical information from a young age. Strong ANS skills are linked to better math performance and overall cognitive development. Although the Subitizing measure captures some of this information, it was not enough in and of itself with such small quantities. The Who Has More measure was chosen to allow children to demonstrate this skill with much larger quantities and varying ratios within the allotted time block. It was also amenable to using features of the iPad as a touch-based measure.

Measure Development

Similar items from the Research-Based Early Math Assessment (REMA) (Clements et al., 2008) were adapted to develop the Who Has More measure, but with more items and images within each item. The REMA presents the child with two pictures of the same items (e.g., fruit) though not of the same quantity (i.e., three and four) and an animal above the respective blocks of items. The administrator asks the child who has more (fruit).

Administration and Scoring

Prior to completion of any touch measure, including Who Has More, the participant should have completed the NIH Baby Toolbox Touchscreen Tutorial. Two arrays are presented on the screen and contain a specified quantity of objects on the left and right sides of the screen. Audio instructions direct the child to choose the set with the larger quantity of images by asking, "Who has more?" The child is prompted to tap the side of the screen containing the set of images that matches their choice.

Items were dichotomous, in which a 0 is given for incorrect responses, and a 1 is given for correct responses. The 22 items were fitted to a unidimensional Rasch (1PL) response model and were scored using ML score estimation.

2.2.14 Verbal Arithmetic

The Verbal Arithmetic measure is administered to children aged 37 months and older and assesses the child's ability to add and subtract in response to verbal word problems.

Measure Development

Verbal arithmetic, the ability to use language to describe and solve numerical problems, plays a significant role in math development. Mastery of these basic operations helps build a strong foundation for understanding more complex mathematical concepts later in life. For infants, engaging in simple addition and subtraction activities enhances cognitive abilities such as problem-solving, logical thinking, and memory.

Administration and Scoring

Simple addition and subtraction questions appear on the iPad; the examiner verbalizes those to the child and then records the child's response on the iPad. Five-word problems are presented in total.

Items were dichotomous, in which a 0 is given for incorrect responses, and a 1 is given for correct responses. Exploratory correlational analyses were conducted to determine the dimensionality of the items in relation to each other. The five items were used to construct confirmatory IRT models, and the data was fitted to a unidimensional Rasch (1PL) response model and were scored using ML score estimation.

2.3 Motor Domain

This section of the chapter describes the goals and objectives, development, and scoring approaches for the measures in the Motor domain of the NIH Baby Toolbox.

Motor measures utilize a similar approach in which the examiner first administers the tasks to the child, recording the task on video through the app. The examiner is then guided through reviewing the video to code components of the child's motor behaviors, including both categorical classifications (e.g., did the child walk or crawl?) as well as quantitative measures (e.g., indicate the times when the child crosses a start/finish line).

2.3.1 Get Up and Go

The Get Up and Go measure assesses postural transition and locomotion in a single task that is appropriate for infants and young children across the entire NIH Baby Toolbox age range (16 days old through 3.5 years old). Two similar versions are administered based on whether the child can locomote, according to the caregiver.

In the Pre-Locomotor version of the measure, children are placed in a supine position and encouraged to roll over into a prone position. In the Locomotor version of the

measure, children are placed in a supine position and encouraged to get up, travel to a caregiver 3 meters away, and get onto and off of a platform.

Measure Development

The Get Up and Go measure is based on the “Timed Up and Go” (TUG) Test (Mathias et al., 1986; Podsiadlo & Richardson, 1991), which has been used extensively with older adult populations. The NIH Baby Toolbox version of Get Up and Go is more developmentally appropriate and offers additional opportunities to assess methods of postural transitions since the child must get out of the supine position (Marsala & VanSant, 1998). Administration and scoring categories were developed and refined following a pilot study with 70 children.

Administration, Observational Coding, and Scoring

For this measure, there are a total of 14 observational coding questions the child may be asked. The exact questions and order of the questions follows branching logic that is based on the responses from prior questions. Responses to individual questions, including the 14 possible observational coding questions and two caregiver questions are then combined into overall rank scores.

This measure produces two subscores: Postural Transition (29 rank scores) and Locomotion (20 rank scores). Responses across the two subscores contribute to the overall Gross Motor score (nine rank scores). Gross Motor scores are calculated by combining Postural Transition and Locomotion scores in a data-driven manner by producing percentile ranks based on data from the norming study.

The following questions are asked of the administrator and caregiver to determine the Postural Transition score:

- Can the child get off their back?
- Did the child get off their back?
- What did the child do on their belly?
- How did the child get off their back?
- What was the most upright posture before the starting line?
- How did the child get into a sitting or kneeling position?
- How did the child get into a standing position?
- How many hands did the child use to get up?

The administrator and caregiver are asked the following questions to determine the Locomotion score:

- Can the child take forward steps independently?
- How did the child travel?

- How far did the child travel?

Additional questions from this measure contribute to the Gross Motor score:

- What did the child do at the stairs?
- How did the child get up the stairs?

Rank scores combine responses across items relevant to that subscore. Children with higher rank scores are considered to exhibit more sophisticated behaviors than those with lower rank scores. The order and comprehensiveness of the ranks for each developmental area were generated and evaluated based on possible and plausible behaviors on the instrument.

A lookup table (Table 2.1) is provided to convert the above Gross Motor rank scores into age-based percentiles. Based on the child's Gross Motor score, their percentile rank compared to all children can be found using the Overall Percentile column in Table 2.1, and their age-based percentile rank with respect to their score can be found by using their age in the Months column.

Table 2.1*Lookup Table to Convert Gross Motor Scores to Percentile Ranks*

Gross motor	Overall percentile	Age in months												
		1-3	4-6	7-9	10-12	13-15	16-18	19-21	22-24	25-27	28-30	31-36	37-42	43-48
1	12%	92%	49%	8%	1%	1%	1%	1%	1%	1%	1%	1%	1%	1%
2	21%	99%	85%	41%	8%	6%	1%	1%	1%	1%	1%	1%	1%	1%
3	31%	99%	99%	79%	53%	24%	9%	3%	1%	1%	1%	1%	1%	1%
4	41%	99%	99%	96%	89%	48%	28%	6%	4%	7%	1%	1%	1%	6%
5	56%	99%	99%	99%	98%	78%	73%	45%	34%	33%	31%	15%	10%	11%
6	69%	99%	99%	99%	99%	95%	93%	70%	68%	61%	45%	34%	21%	19%
7	81%	99%	99%	99%	99%	98%	95%	80%	82%	77%	70%	63%	52%	43%
8	89%	99%	99%	99%	99%	99%	99%	93%	94%	93%	83%	75%	70%	54%
9	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%	99%

2.3.2 Reach to Eat

The Reach to Eat measure assesses a child's reaching, grasping, and tool-use behaviors in a single measure that is appropriate for infants and young children across the entire NIH Baby Toolbox age range (16 days old through 3.5 years old). Two similar versions are administered based on whether the child is older than 6 months old and whether their caregiver is comfortable with them eating solid foods (i.e., a Cheerio).

In the standard version (for children over 6 months old who can eat Cheerios), the child is presented with a Cheerio balanced on a small base. The child is encouraged to grasp the Cheerio across two trials, one using their right hand and one using their left. Grasping the Cheerio directly from the base without it falling to the table is considered a "success." If the child cannot succeed with either hand, they complete the same task on a larger-sized base.

If the child succeeds in lifting the Cheerio from the small base on at least one trial (with either hand), the child proceeds to spoon trials. For spoon trials, a Cheerio is placed on a spoon that sits inside a cup. The handle of the spoon points either in the direction of the child's target hand (easy trials) or away from the child's target hand (hard trials). The child is encouraged to grasp the spoon in order to transport the Cheerio to their mouth.

Children younger than 6 months old and/or who are not eating solid foods are administered the Block version of the measure. A 1-inch block is placed on the large base. Children are encouraged to reach and grasp the block across two trials—one with each hand.

Measure Development

The concept of the Reach to Eat (Cheerio) version of the measure is adapted from class work on infant reaching behavior (Sacrey et al., 2012), whereas spoon trials were developed based on work examining tool-use and manual planning (Keen et al., 2014; McCarty et al., 1999, 2001). The Reach to Eat (Block) version of the measure is an adaptation of work by Gesell (1926).

Administration, Observational Coding, and Scoring

Examiners administer the measure while recording the event on the iPad. The recorded video is then available for review for the examiner to complete observational coding.

There are a maximum of 32 observational coding questions (16 per hand). The questions asked and the order of the questions follow branching logic that is based on the responses from prior questions. Responses to individual questions are then combined into overall rank scores.

This measure produces either a ranked Manual Dexterity (Block) score or a combination of the following IRT-based scores based on completion of the Cheerio version of the measure. Those who complete the small base and spoon portions of the Cheerio version of the measure may receive a Manual Dexterity, Manual Planning, and Fine Motor score, while those who complete the small base and large base portions of the Cheerio version of the measure will receive only a Manual Dexterity and Fine Motor score.

The Manual Dexterity score utilizes items relating to whether and how the child successfully grasped the Cheerio from the small base, large base, and/or using the spoon, as well as the type of grasp they used (Success and Grasp related items, see item-level data dictionary). Of note, no Manual Dexterity Score will be produced if all relevant Success and Grasp items related to the spoon portion are missing, which can occur if Purpose or Success related items for the spoon portion are marked with responses related to measure noncompliance (see Administration Manual for further details about the measure's branching logic and criteria for noncompliance). Items for the Manual Dexterity score were calibrated using a GRM.

The Manual Planning score utilizes items relating to whether the child's thumb was pointing toward or away from the bowl while holding the spoon and whether the child grasped the spoon handle or moved before grasping (Move and Thumb related items, see item-level data dictionary). These items can be completed only if the child grasped the spoon with the purpose of transporting the Cheerio (see Purpose related items in the item-level data dictionary). Items for the Manual Planning score were calibrated using a 2PL IRT model. Examinees were scored using maximum likelihood scoring to obtain an interval-scaled score on the logit metric.

The Fine Motor score combines items used towards the Manual Dexterity and Manual Planning (if available) scores. The Fine Motor score uses mixed-item types (dichotomous items from Manual Planning calibrated using 2PL and polytomous items from Manual Dexterity calibrated using a GRM). As such, Fine Motor scores are obtained using the MAP calculator (Section 2.1). All Manual Dexterity, Manual Planning, and Fine Motor scores produced by the Reach to Eat (Cheerio) version of the measure are converted to a CSS for reporting purposes. Notably, the Fine Motor score can be produced with responses relevant to the Manual Dexterity score alone or from both Manual Dexterity and Manual Planning. However, a Fine Motor score will not be produced with Manual Planning alone.

The Manual Dexterity (Block) scores for the Right Hand (RH) and Left Hand (LH) are ranked based on responses to the questions about whether the child reached with their right hand (RH)/left hand (LH): "Did the child reach for the block with their right/left hand?" and "What type of grasp did they use (palmar, multi-finger, thumb-and-finger)?"

A Manual Dexterity Average score of the mean of the RH and LH scores are also provided. The ranks below describe the Dexterity scores (see Table 2.2).

Table 2.2

Manual Dexterity Scores for the Reach to Eat (Block) Version of the Measure for both Right Hand (RH) and Left Hand (LH)

Manual dexterity Reach to Eat (Block) RH and LH rank scores	Description
1	Didn't try
2	Moved arm only
3	Touched but no grasp
4	Grasped from table + used a palmar, multi-finger, or thumb-finger grasp type
5	Grasped from base + palmar grasp
6	Grasped from base + multi-finger grasp
7	Grasped from base + thumb and finger grasp

2.4. Social-Emotional Functioning Domain

This chapter describes the development and scoring approaches for the measures in the Social-Emotional Functioning domain of the NIH Baby Toolbox. There are six subdomains, as described in Table 2.3. Measures include caregiver report and direct observation with live examiner coding.

Table 2.3

Social-Emotional Functioning Subdomains and Definitions

Subdomain	Definition
Social Communication	Skills involving joint attention, recognizing and responding to others' communicative behaviors, sharing one's own intentions in communication, and using the intentions of others to guide one's own behavior. Social communication also includes following conversational rules that relate to initiating and maintaining reciprocity in conversations, repairing miscommunications, following communicative conventions, and accounting for the perspective of others in social interactions and communication.
Social Relationships	Positive relationships, experiences, and connectedness with caregivers, peers, and adults; sociability; and the engagement in empathic behaviors (Blackwell, Lai, et al., 2022).
Temperament	Biologically-based individual differences in emotion and emotional reactivity and regulation (Aktar & Pérez-Edgar, 2020).
Negative Affect	Negative Affect refers to unpleasant feelings or emotional distress related to early internalizing (e.g., sadness, fear) and externalizing (e.g., anger, irritability) behaviors (Sherlock et al., 2022) on a continuum from normal expressions sadness, fear, and anger/irritability to more extreme expressions.

Psychological Well-Being	Positive or rewarding affective experiences and engagement with the environment, such as happiness, joy, enthusiasm, and excitement (Salsman et al., 2014).
Self-Regulation	Ability to effectively recognize and control emotions and behaviors in the service of one's own goals and adapt in response to environmental demands and expectations (Blackwell, Lai, et al., 2022).

2.4.1 Caregiver Checklist

The Caregiver Checklist measure evaluates children's social-communication skills. There are two versions of the Caregiver Checklist measure, one for younger children aged 6-23 months and one for older children aged 24-42 months. Within each version, there are age-based item sets, where each form is tailored to a specific age band within each version's age range (i.e., 6-11, 12-17, 18-23 months old; 24-29, 30-35, 36-41, 42 months old). Scores within the broader age categories (6-23 months old and 24-42 months old) are comparable with each other, though the item sets include unique items by the forms. The Caregiver Checklist measure provides a broad assessment of children's interactions with others and their social environment that is not feasible to assess in a structured observation.

Measure Development

The Caregiver Checklist measure was developed in part based on existing caregiver report checklists related to social communication (Wetherby et al., 2008). The versions of the measure for younger children (6-23 months old) include items from the Communication and Symbolic Behavior Scales Developmental Profile™ Infant-Toddler Checklist (used with permission from Wetherby et al., 2008). Items in the versions of the measure for older children were developed *de novo* to conceptually align to the areas assessed in the 6-23 months old versions of the measure. The 42-month-old Caregiver Checklist measure is intended for children aged 42 months but contains age-appropriate content for children up to 48 months old.

Administration and Scoring

The items within each Caregiver Checklist measure version (6-23 months old; 24-42 months old) are comprised of four item sets (Joint Attention/Shared Attention; Play; Prosocial; Social Communication), which are considered when determining whether the instrument is eligible to be scored. Several additional items in each version of the measure are not part of any item set. The measure-level IRT score is calculated from a bifactor graded response model, where only the general factor is reported. Although each measure contains multiple forms (6-11, 12-17, 18-23 months old; 24-29, 30-35, 36-41, 42 months old), vertical scaling was used to permit scoring each version on a common scale. For each version, the IRT score and standard error are calculated when **all** the following conditions are met:

- More than 50% of items in the entire checklist have a response recorded.

- More than 50% of the items within each item set have a response recorded.

Here, “entire checklist” and “each item set” refer only to items in the form administered; for example, for a caregiver completing the form for children aged 6-11 months, only items on that form are used for the eligibility rules above. Items without an item set contribute to the first eligibility criterion but not the second. Bifactor IRT scoring (Reise, 2012) also yields group factor scores for items in each item set after accounting for variance explained by the general (test-wide, reported) latent variable; this definition of group factors as “residual” to the general factor renders these group factors difficult to interpret, and they are therefore not reported.

2.4.2 Social Observation Measure (SOM)

The Social Observation Measure (SOM) assesses children’s social functioning in a semi-structured administrator–led interaction and in a less structured caregiver-child interaction. There are two age-based versions of the SOM: the SOM–Younger version for children aged 9-23 months and the SOM–Older version for children aged 24-42 months.

Measure Development

The SOM development was informed by existing, well-validated measures, such as the Communication and Symbolic Behavior Scales (Wetherby et al., 2008) and utilized a series of iterative evaluations to produce the final measure. Initial draft versions of the SOM were developed and piloted for use with children aged 6-23 months and 24-42 months. The pilot study evaluated the feasibility of administration and the scoring and effectiveness of training procedures. Based on the quantitative data (e.g., fidelity of administration; scoring accuracy) collected and qualitative feedback from examiners, administration and scoring were simplified and training materials and procedures were revised. The revised SOM underwent a second round of testing to evaluate feasibility and to collect preliminary validity data. Results were used to further improve administration, scoring and training procedures, and indicated that the SOM was not suitable for use with very young children. Thus, the final SOM–Younger version begins at 9 months of age instead of the originally intended 6 months of age.

Administration and Scoring

Both SOM versions are structured in 1- or 2-minute intervals that, together, yield a 10-minute assessment. These intervals were designed to encourage social communication and reciprocal social interaction, and enable the assessment of social attention, social communication, play, and prosocial behavior. Each interval involves administrator prompts and coding child behavior ratings. All ratings require a response from the examiner to proceed.

2.4.2.1 SOM–Younger

The app combines the interval-level ratings coded by the administrator into items for scoring. For the SOM–Younger version, these items are grouped into four item sets: Shares Attention, Shares Enjoyment, Shares Interest, and Play. Each item set yields two scores: (1) a summed score of all items comprising that item set; and (2) an IRT score, and standard error calculated using a multidimensional GRM, which incorporates scores from other item sets to improve measurement precision. In addition, scores are reported representing a child’s ability to respond to their own name (SOM Responds to Name Score) and follow another’s gaze (SOM Follows Gaze Score), calculated as the sums of the corresponding items across intervals.

2.4.2.2 SOM–Older

For the SOM–Older version, the app similarly combines the interval-level ratings coded by the administrator into items for scoring; however, unlike the SOM–Younger version, there are no item sets. The SOM–Older version yields one summed score and one IRT score, and standard error calculated using the unidimensional graded response model.

2.4.3 Infant, Early Childhood, and Childhood Behavior Questionnaires

There are three age-based caregiver report questionnaires that assess the Temperament subdomain of the Social-Emotional Functioning domain. The Infant Behavior Questionnaire–Revised Very Short Form (IBQ-R-VSF; Putnam et al., 2014) is appropriate for 3- to 12-month-olds; the Early Childhood Behavior Questionnaire–Very Short Form (ECBQ-VSF; Putnam et al., 2006) is appropriate for 13- to 36-month-olds; and the Childhood Behavior Questionnaire–Very Short Form (CBQ-VSF; Putnam & Rothbart, 2006) is appropriate for 37- to 42-month-olds.

Measure Development

The IBQ-R-VSF (Putnam et al., 2014), ECBQ-VSF (Putnam et al., 2006), and CBQ-VSF (Putnam & Rothbart, 2006) are existing, widely used validated caregiver-report measures of an infant and young child’s temperament and were integrated with the NIH Baby Toolbox with permission. They were developed as parallel age-based measures, with conceptual and structural coherence, each with three subscales for effortful control, surgency, and negative affect. They were originally developed using classical test theory without national norms. The NIH Baby Toolbox team applied IRT methods and normed to the U.S. population.

Administration and Scoring

2.4.3.1 Infant Behavior Questionnaire–Revised Very Short Form (IBQ-R-VSF)

The IBQ-R-VSF (Putnam et al., 2014) is completed by caregivers of children aged 3-12 months on an iPad. First, parents are presented with an instructional screen, followed by one item per screen. Items are administered sequentially. Once a caregiver responds to an item, the system will auto advance to the next item. If a caregiver does not tap a response and instead taps “Skip,” an alert will appear. The caregiver must confirm they want to skip the item by tapping “Skip” or cancel the skip by tapping “Cancel.” Items ask the caregiver about the infant’s behavior over the last week on a 7-point Likert scale from Never to Always. A “Does Not Apply” option is also available if the parent has not seen the infant in the situation in the past week.

The Baby Toolbox IBQ-R-VSF yields four scores, three of which reflect the original Putnam et al. (2014) subscales (Surgency, Negative Affect, and Effortful Control). The fourth score describes Fear of Unfamiliar Adults (“Stranger Danger”) and is derived from a subset of three items from Negative Affect; these items are also used in the calculation of Negative Affect scores. Of note, the original IBQ-R-VSF used the term “Orienting/Regulatory Capacity” to describe the item set that NIH Baby Toolbox has termed “Effortful Control.” This decision was made to enable consistency across versions and because the Orienting/Regulatory Capacity subscale is a known predictor of effortful control at later ages (Putnam et al., 2008).

For each item set, scores are calculated if more than 50% of items within a set have recorded responses other than “Does Not Apply.” Eligibility is assessed per item set, (i.e., if only one item set has 50% or fewer responses, instrument scores will be produced for the remaining item sets). Two scores are produced for each item set: a mean score, calculated as the mean of recorded responses for items in that set; and an IRT score calculated using the MAP calculator (Section 2.1), which incorporates scores from other item sets to improve measurement precision. To facilitate IRT scoring, some score options for each item are collapsed to a single option prior to applying the MAP calculator.

Fear of Unfamiliar Adults (“Stranger Danger”) Score: When developing scoring models, three items representing fear of unfamiliar adults were disproportionately weighted on the factor corresponding to Negative Affect; further examination motivated scoring these items separately on a factor representing Fear of Unfamiliar Adults (or “Stranger Danger”). Thus, this score is not part of the original IBQ-R-VSF scoring and is new for the NIH Baby Toolbox.

2.4.3.2 Early Childhood Behavior Questionnaire–Very Short Form (ECBQ-VSF)

The ECBQ-VSF (Putnam et al., 2006) is completed by caregivers of children aged 13-36 months on an iPad. Caregivers are first presented with an instructional screen

describing the questionnaire, followed by one item per screen. Items ask about the child's behavior over the last 2 weeks on a 7-point Likert scale from Never to Always. A "Does Not Apply" option is also available if the caregiver has not seen the child in the situation in the past 2 weeks.

The ECBQ-VSF yields three scores based on three item sets: Surgency, Negative Affect, and Effortful Control. For each item set, scores are calculated if more than 50% of items within the set have recorded responses other than "Does Not Apply." Eligibility is assessed per item set (i.e., if only one set has 50% responses or fewer, instrument scores will be produced for the remaining item sets). Two scores are produced for each item set: a mean score, calculated as the mean of recorded responses for items in that set; and an IRT score calculated using the MAP calculator (Section 2.1), which incorporates scores from other item sets to improve measurement precision. To facilitate IRT scoring, some score options for each item are collapsed to a single option prior to applying the MAP calculator.

2.4.4.3 Childhood Behavior Questionnaire–Very Short Form (CBQ-VSF)

The CBQ-VSF (Putnam & Rothbart, 2006) is completed by caregivers of children aged 37-42 months on an iPad. First, caregivers are presented with an instructional screen describing the questionnaire, followed by one item per screen. Items ask about the child's behavior over the last 6 months on a 7-point Likert scale from Extremely Untrue to Extremely True. A "Does Not Apply" option is also available if the caregiver has not seen the child in the situation in the past 6 months.

The CBQ-VSF yields three scores based on three item sets: Surgency, Negative Affect, and Effortful Control. For each item set, scores are calculated if more than 50% of items within the set have recorded responses other than "Does Not Apply." Eligibility is assessed per item set (i.e., if only one set has 50% responses or fewer, instrument scores will be produced for the remaining item sets). Two scores are produced for each item set: a mean score, calculated as the mean of recorded responses for items in that set; and an IRT score, calculated using the MAP calculator (Section 2.1), which incorporates scores from other item sets to improve measurement precision. To facilitate IRT scoring, some score options for each item are collapsed to a single option prior to applying the MAP calculator.

2.4.4 Patient-Reported Outcomes Measurement Information System (PROMIS®) Early Childhood (EC) Parent Report Measures

The Social-Emotional Functioning domain includes eight PROMIS EC parent-report measures outlined by subdomains in Table 2.4, adapted with permission for the NIH Baby Toolbox.

Table 2.4*PROMIS EC Measures by Subdomain*

Subdomain	PROMIS EC measure
Negative Affect	PROMIS EC Parent-Report Bank v1.0 - Anger/Irritability
	PROMIS EC Parent-Report Bank v1.0 - Anxiety
	PROMIS EC Parent-Report Bank v1.0 - Depressive Symptoms
Psychological Well-Being	PROMIS EC Parent-Report Bank v1.0 - Positive Affect
Self-Regulation	PROMIS EC Parent-Report Scale v1.0 - Self-Reg - Frustration Tolerance 6a
	PROMIS EC Parent-Report Scale v1.0 - Self-Reg - Flexibility 5a
Social Relationships	PROMIS EC Parent-Report SF v1.0 - Soc Rel - Child-Caregiver Interactions
	PROMIS EC Parent-Report SF v1.0 - Soc Rel - Peer Relationships 4a

Measure Development

PROMIS EC instruments are brief (4-6 items or computerized adaptive testing [CAT]) parent- or caregiver-reported surveys developed using state-of-the-science PROMIS mixed methods methodology that included in-depth literature reviews; interviews with experts and caregivers to elicit key concepts to assess; caregiver cognitive interviews to assess readability and comprehensibility; large-scale field testing with over 2,000 U.S. caregivers of children aged 1-5 years; and the application of IRT for psychometric validation. For an overview of the PROMIS EC methodology, see Cella et al. (2022) and Lai et al. (2022).

All PROMIS EC measures are normed to a general U.S. population of caregivers of children aged 1-5 years. Descriptions of the development process and psychometric validation for individual instruments can be found elsewhere (Anger/Irritability, Anxiety, Depressive Symptoms (Sherlock et al., 2022); Positive Affect, Flexibility, Frustration Tolerance (Blackwell, Kallen, et al., 2022); Child-Caregiver Interactions, Peer Relationships (Blackwell, Lai, et al., 2022). PROMIS EC measures were adopted by the NIH Baby Toolbox with permission from the PROMIS Health Organization.

Administration and Scoring

PROMIS EC measures are completed by caregivers of children aged 13-42 months on an iPad. Caregivers are presented with one item on a screen at a time with five response options ranging from Never to Always or Never to Almost Always. Once a caregiver responds to the item on the screen, the system will auto advance to the next item. If a caregiver does not tap a response and instead taps “Skip,” an alert will

appear. The caregiver must confirm they want to skip the item by tapping “Skip” or cancel the skip by tapping “Cancel.”

PROMIS EC Scales (i.e., Flexibility, Frustration Tolerance, Child-Caregiver Interactions, Peer Relationships) are administered as static forms where all caregivers receive the same set of 4-6 items per measure.

PROMIS EC Item Banks (i.e., Anger/Irritability, Anxiety, Depressive Symptoms, Positive Affect) are administered based upon an internally developed CAT algorithm. The CAT algorithm selects items for administration, such that only a subset of item bank items is administered. The algorithm only selects items that sharpen the estimate of a respondent’s score on the domain being measured. Stopping rules are applied to inform the system when enough information is available to provide instrument-level scores.

All PROMIS EC measures are scored using an IRT method to obtain scores on the PROMIS T score metric (mean = 50, standard deviation [SD] = 10). Scoring is calculated directly within the NIH Baby Toolbox app. Higher scores reflect more of the concept being measured (e.g., More Anger/Irritability, More Positive Affect).

2.5 Experimental/Supplemental Measures

2.5.1 Sit and Stand

As a supplemental motor measure, children’s stationary postural control is assessed using the Sit and Stand measure. Stationary postural control in sitting and standing are classic measures of motor development because they reflect children’s increasing control over the segments of their spine (McGraw, 1940). The NIH Baby Toolbox Sit and Stand measure is adapted from the NIHTB Standing Balance Test (Rine et al., 2013). Based on the child’s age and caregiver questions related to the child’s ability to sit, stand, and walk in an unsupported fashion, the child is administered certain tasks assessing their ability within what is developmentally appropriate.

Administration and Scoring

Children who can walk and are older than 24 months old are asked to hold two postures for up to 30 seconds: feet together and feet tandem. Children who can stand but not walk are encouraged to hold one position for up to 30 seconds: feet apart stand. Children who can sit but cannot yet stand complete two tasks: pull-to-sit and legs together sit. Walkers who are less than 24 months old are not tested because pilot testing indicated that they would not comply with instructions to stand still.

The length of time that children can maintain balance in the specified posture reveals their level of balance control. Examiners also indicate whether the child was able to maintain the posture for the full 30 seconds or whether they fell, stepped out, or walked/crawled away.

Here, Stationary Postural Control is assessed via a rank-based score, like in the Get Up and Go measure (see Table 2.5).

Table 2.5

Sit and Stand Measure Stationary Postural Control Rank Scores

Stationary postural control	Description
1	Head lag in pull-to-sit
2	No head lag in pull-to-sit
3	Fell from unsupported sit
4	Used hands for support in sit
5	Sat 30 seconds without support in sit
6	Shifted from sit
7	Feet apart fell or grabbed
8	Feet apart shifted to floor
9	Feet apart stepped out
10	Feet apart stood 30 seconds or caregiver said child can walk 3 meters but are younger than 24 months
11	Feet together fell or grabbed
12	Feet together shifted to floor
13	Feet together stepped out
14	Feet together stood feet together 30 seconds
15	Feet tandem fell or grabbed
16	Feet tandem shifted to floor
17	Feet tandem stepped out
18	Feet tandem stood feet tandem 30 seconds

Given the difficulties with administration compliance, data from the norming study were deemed unacceptable and the measure was moved to supplementary. It is recommended primarily for examiners with experience administering motor-related measures.

3. Norming of the NIH Baby Toolbox

The NIH Baby Toolbox norming study data collection was conducted by a private market research firm under the direction of researchers at Northwestern University. The firm hired examiners, recruited participants, and conducted assessments at their offices. Norming testing for the NIH Baby Toolbox began in spring 2023 across 16 sites, representing all nine divisions and four regions in the U.S. Census Bureau, and concluded in March 2024.

Table 3.1

All Sites and Regions in which NIH Baby Toolbox Testing Occurred

Region	Sites
Midwest	Appleton (WI), Chicago (IL), Minneapolis (MN), St. Louis (MO)
Northeast	Boston (MA), Iselin (NJ), Philadelphia (PA)
South	Atlanta (GA), Baltimore (MD), Dallas (TX), Houston (TX), Nashville (TN), Orlando (FL)
West	Los Angeles (CA), Phoenix (AZ), Seattle (WA)

3.1 Description of the Norming Study

The marketing research firm staff were trained to administer the NIH Baby Toolbox using a Train-the-Trainer model. Certified trainers were taught to administer and score the NIH Baby Toolbox by the centralized training team at Northwestern University. Two certified trainers per site (30 trainers total) were required to attend a 1-week training workshop, practice administration with both adults and children, submit two recorded administrations (one with an infant aged 0-24 months, one with a toddler aged 24-48 months), and submit observational scoring of eight videos. Trainees received comprehensive feedback regarding all aspects of administration and were asked to submit new videos if more than 75% errors were made during administration steps or were asked to score additional videos if they achieved lower than 75% accuracy on scoring. Certified trainers then used the same training materials to conduct their own training workshops at each site. Depending on the site volume, 5-15 staff were trained by the certified trainer team. Trainers completed the first level of certification for site administrators by using checklists and writing summaries to provide feedback on practice cases. The NIH Baby Toolbox training staff then reviewed the feedback for all 90 testers, provided additional comments, and determined whether a given tester achieved adequate fidelity in item administration and scoring for items not automatically scored by the test administration software.

Children were tested together with their caregivers present at one of the 16 market research firm locations. Following a warm-up period that lasted at least 5-10 minutes,

caregivers were given an iPad to complete caregiver-report questionnaires while the child was administered the direct assessment portion of the NIH Baby Toolbox battery based on their age. See Table 3.2 for the intended battery order for each age band. Examiners were allowed to skip or delay-and-resume measures if indicated based on rapport or child need during the assessment. All caregivers were also asked to complete a sociodemographic questionnaire that asked additional questions about social determinates of health, child and family health characteristics, media use, and other relevant clinical and demographic data.

Table 3.2*Intended Battery Order of the NIH Baby Toolbox Norming Study by Age Bands in Months*

Order	1-5 months	6-8 months	9-15 months	16-21 months	22-24 months	25-36 months	37+ months
1	Sit and Stand	NBT Mullen Visual Reception	Social Observation	Social Observation	Social Observation	Social Observation	Social Observation
2	Get Up and Go	Sit and Stand	Executive Function ¹	Executive Function ¹	Mullen Receptive	NBT Touch Screen Tutorial	NBT Touch Screen Tutorial
3	Mullen Receptive	Get Up and Go	Mullen Receptive	Mullen Receptive	Mullen Expressive Prompted	Memory Task (Learning)	Memory Task (Learning)
4	Mullen Expressive Prompted	Executive Function	Looking While Listening	Looking While Listening	Looking While Listening	Visual Delayed Response	Visual Delayed Response
5	Mullen Visual Reception	Mullen Receptive	Mullen Expressive Prompted	Mullen Expressive Prompted	Sit and Stand	Memory Task (Test)	Memory Task (Test)
6	Reach to Eat	Looking While Listening	Numerical Change Detection*	Numerical Change Detection*	Get Up and Go	Mullen Receptive	Mullen Receptive
7	Mullen Expressive Observational	NBT Mullen Expressive Prompted	Sit and Stand	Sit and Stand	Numerical Change Detection*	Mullen Expressive Prompted	Mullen Expressive Prompted
8		Numerical Change Detection*	Get Up and Go	Get Up and Go	Spatial Change Detection*	Picture Vocabulary	Picture Vocabulary
9		Reach to Eat	NBT Mullen Visual Reception	NBT Mullen Visual Reception	NBT Touch Screen Tutorial	Sit and Stand	Sit and Stand
10		Spatial Change Detection*	Spatial Change Detection*	Spatial Change Detection*	Memory Task (Learning)	Get Up and Go	Get Up and Go
11		Mullen Expressive Observational	Reach to Eat	Reach to Eat	Visual Delayed Response	Verbal Counting	Verbal Counting
12			Mullen Expressive Observational	Mullen Expressive Observational	Memory Task (Test)	Object Counting	Object Counting
13					Mullen Visual Reception	Subitizing	Subitizing
14					Reach to Eat	Who Has More	Who Has More
15					Mullen Expressive Observational	Mullen Visual Reception	Verbal Arithmetic
16						Reach to Eat	Mullen Visual Reception
17						Mullen Expressive Observational	Reach to Eat
18							Picture Sequence Memory*
19							Speeded Matching*
20							Mullen Expressive Observational

Note. The measures with an asterisk (*) are not part of the NIH Baby Toolbox but were administered during the norming study.¹ The Executive Function battery was renamed to Familiarization post-norming.

3.1.1 Characteristics of the Norming Sample

Strata defining relevant demographic groups were proposed during the creation of the sampling plan prior to data collection. The primary demographic variable of interest was chronological or adjusted age in months (CAAMs). At the time of the norming study, adjusted age accounted for premature birth if the child's chronological age was 36 months old or younger and if the child was born at least 3 weeks prior to the expected due date. Following best practices in continuous norm derivation, target sample sizes for age were based on score precision (as opposed to population representation). Age was rounded to the nearest integer month (16 days old rounded up, 15 days old rounded down). Strata were in 1-month increments from 1-6 CAAMs (6 strata), 2-month increments from 7-12 CAAMs (3 strata), 3-month increments from 13-30 CAAMs (6 strata), and 6-month increments from 31-48 CAAMs (3 strata). Due to practical constraints recruiting infants, the target sample size was 100 per strata for 1-3 CAAMs and 150 per strata for 4 CAAMs and above.

Other relevant demographic information was sampled proportionate to distributions from the U.S. Census Bureau, specifically the 2019 American Community Survey (ACS) Public Use Microdata Sample (PUMS ACS; U.S. Census Bureau, 2019). Demographic targets were nested within geographic region, recognizing that diversity and representation varies across the country. Relevant demographic variables included race and ethnicity by language exposure, sex assigned at birth, and parental education. Targets for sex assigned at birth were nested within age strata. Parental education and the cross between race/ethnicity and language were nested in broader age categories to prevent small (and thus unattainable) target sample sizes. Among Spanish-speaking households, targets were based on Hispanic or Latino ethnicity for 1-6 CAAMs, 7-12 CAAMs, 13-24 CAAMs, 25-36 CAAMs, and 37-48 CAAMs. For Spanish-speaking households that were non-Hispanic, targets were based on any age, geographic region, and racial identity. For English-speaking households, targets were based on Hispanic or Latino ethnicity (regardless of racial identity), or Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian, and Non-Hispanic Other (including multiracial) for the same CAAM age strata as Spanish-speaking Hispanic individuals (1-6, 7-12, 13-24, 25-36, and 37-48 CAAMs). These strata were nested within geographic region, except for Non-Hispanic Asian, which could be recruited regardless of geographic region given census proportions of primarily English- or Spanish-speaking households in the target age range. Parental education targets were based on the same broader age strata nested within geographic region for High School Graduate, GED, or Less; Some College (including 2-year associate's degree and technical school degree); Bachelor's degree; and Advanced degree (master's, doctorate, or professional degrees).

A 10% margin around the 2019 5-year ACS proportions were considered acceptable for the purpose of sampling during the data collection phase of the study. After data

collection was complete, the sampling cells were updated from the 2019 5-year ACS estimates to the 2022 5-year ACS estimates (U.S. Census Bureau, 2022). These updated proportions were then used to compute the examinee weights for the NIH Baby Toolbox norm calculations. Figure 3.1 provides a density plot for the sample weights by broad age group. By construction, the mean sample weight was 1.0. The largest sample weights were assigned to individuals from demographic groups less-represented in the sample, particularly Spanish-speaking children of parents with less than a high school education.

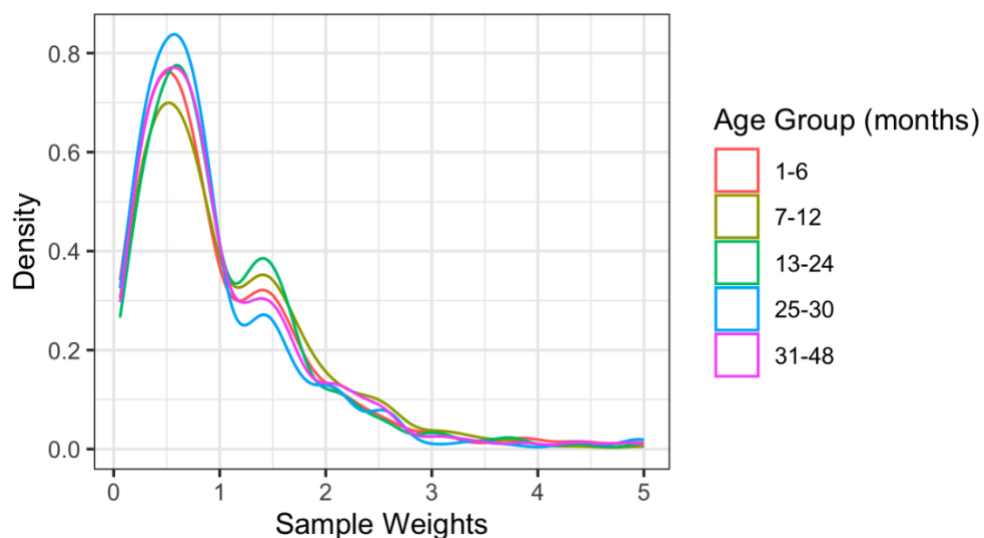


Figure 3.1 *Distribution of Sampling Weights from the NIH Baby Toolbox Norming Study*

The sample demographics for children aged 1-6, 7-12, 13-24, 25-30, and 31-48 months, and the total sample, are reported in Table 3.3. As appropriate, the demographic categories were aggregated and combined for weighting purposes. The demographic unweighted and weighted distributions for the variables of interest are also reported in Tables 3.3. The raw count of participants is provided, as well as the population proportion based on the 2022 5-year ACS. The unweighted and weighted proportions for the obtained sample are then reported, with comparisons to the target population proportions.

Table 3.3*U.S. Census Bureau Representation of the NIH Baby Toolbox Norming Sample and Norm Weights Table*

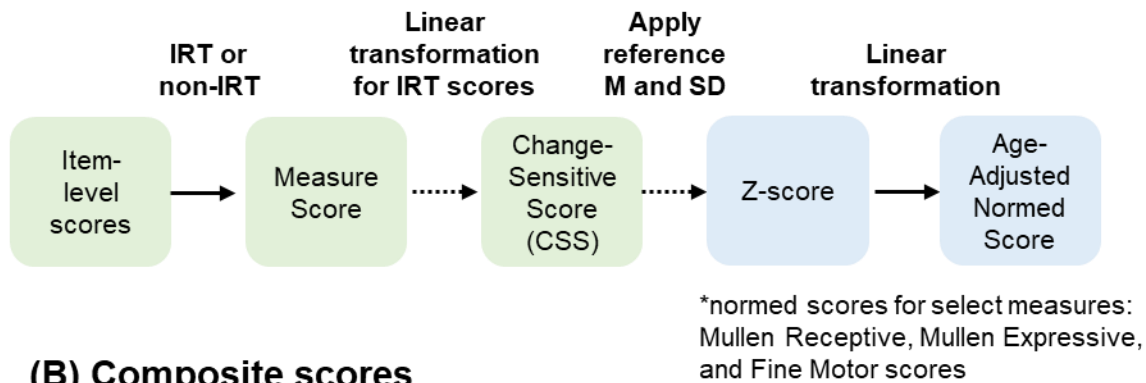
Demographic	Category	Target	N	Unweighted %	Weighted %	Weighted difference¹
Geographic Region	Northeast	15.9%	291	11.6%	15.9%	0.0%
	Midwest	21.4%	814	32.4%	21.4%	0.0%
	South	39.1%	981	39.0%	39.0%	0.0%
	West	23.7%	429	17.1%	23.7%	0.0%
Primary Language	English	76.5%	2,042	81.2%	76.5%	0.0%
	Spanish	23.5%	473	18.8%	23.5%	0.0%
Sex Assigned at Birth	Female	48.9%	1,236	49.1%	50.4%	1.5%
	Male	51.1%	1,279	50.9%	49.6%	1.5%
Race/Ethnicity	Hispanic (any Race)	26.0%	856	34.0%	26.0%	0.0%
	Non-Hispanic White	47.5%	835	33.2%	47.5%	0.0%
	Non-Hispanic Black	12.9%	536	21.3%	12.9%	0.0%
	Non-Hispanic Asian	5.0%	55	2.2%	5.0%	0.0%
	Non-Hispanic Other	8.6%	233	9.3%	8.6%	0.0%
Parental Education	Less than High School Graduate	10.4%	114	4.5%	10.4%	0.0%
	High School Graduate or GED	23.2%	515	20.5%	23.2%	0.0%
	Some College (Including Associate's or Technical Degree)	29.6%	759	30.2%	29.6%	0.0%
	Bachelor's Degree	22.6%	700	27.8%	22.6%	0.0%
	Advanced Degree	14.2%	427	17.0%	14.2%	0.0%

¹ Difference in weighted percentage and target demographic percentage.

3.2 Norming Procedures

For the following procedures outlined by DeBell and Krosnick (2009), iterative proportional fitting was used to derive probability weights for each individual, with a maximum weight of 5. Following Oosterhuis et al. (2016), bootstrap resamples were regressed on age in months to develop the age-adjusted norms for each measure.

(A) Individual Measure Scores



(B) Composite scores

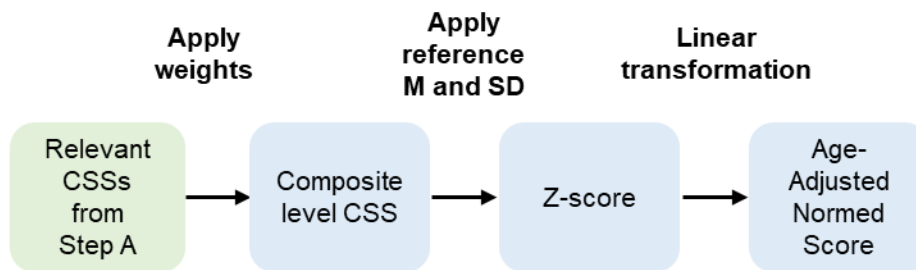


Figure 3.2 Procedure for Norming Measure and Composite Scores for the NIH Baby Toolbox

3.2.1 Norming of Individual Test and Composite Scores

The Mullen Receptive, Mullen Expressive, and Fine Motor (Reach to Eat) measure-level scores were normed. Additionally, Picture Vocabulary measure scores used norms derived from the NIHTB for CAAMs aged 36-42 months, see the V3 Technical Manual (LaForte et al., 2024). Prior to norming, the final thetas and corresponding standard errors were derived from IRT scoring and were converted to CSSs and CSS standard errors through a linear transformation, such that scores became positive integers (see Figure 3.2).

Composite scores for Executive Function/Memory, Language, Motor, Cognition, Self-Regulation, Negative Affect, and Social Communication domains and subdomains were

also normed. The raw composite scores, which were the weighted sum of specific measure-level scores as derived from factor analyses (Chapter 4), were also converted to CSSs prior to norming procedures.

Individual sampling weights were applied to every case in the sample. Bootstrap resamples were drawn and each measure-level or composite-level CSS was utilized for the selected cases. Within each resample, plausible value imputation was used to sample an expected score for each examinee based off of their obtained CSS and its associated standard error. The bootstrap resamples were regressed on age using Generalized Additive Models for Location, Scale, and Shape (GAMLSS) to develop age-adjusted norms for each measure (Stasinopoulos et al., 2018). For the norming, only location and scale modeling were included, both of which were regressed on participant-level CAAMs. Scores greater than 146 or less than 54 were winsorized to 146 or 54, respectively, which were 1 point more extreme than ± 3 SD above or below the mean. One exception is the Picture Vocabulary measure, which follows scoring rules from the original NIHTB, see the V3 Technical Manual (LaForte et al., 2024). The age-adjusted norm scores for the Picture Vocabulary measure were distributed at $M = 100$ and $SD = 15$ for children aged 36 months and older only.

4. Reliability and Validity of the NIH Baby Toolbox

4.1 Reliability

Reliability generally refers to the stability of test scores and includes many types; below are IRT-based reliability and test-retest reliability for the NIH Baby Toolbox (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 2014). For both, high reliability indicates that score differences reflect genuine changes in the latent trait, or the underlying variable being measured.

4.1.1 Empirical Reliability

For all NIH Baby Toolbox measure scores derived using IRT, empirical reliability was calculated for CSS using the following Equation 4.1 (Maydeu-Olivares & Brown, 2010):

$$\text{Equation 4.1: } CSS_{R11} = \frac{\text{Variance}_{CSS}}{(\text{Variance}_{CSS} + \text{mean}(\text{SEM}_{CSS}^2))}$$

applied to the appropriate measures and/or specific age bands within measures for all norming participants. Empirical reliability reflects the variance attributable to true differences in the latent trait, rather than to random measurement error. Sample sizes, reliability coefficients, and standard errors of measurement (SEM CSS) for all Cognition domain measures are presented in their domain-specific tables (see Tables 4.1-4.4).

SEM (CSS) is based on the conditional standard error of the CSS calculated while accounting for measure or score-level empirical reliability, following Equation 4.2:

$$\text{Equation 4.2: } SEM(CSS) = \frac{SD(CSS)}{\sqrt{1 - CSS_{R11}}}$$

4.1.2 Cognition Domain Reliability

Overall, measures within the Cognition domain (Language, Executive Function/Memory, Math subdomains) showed moderate to high reliability. Language domains, except for the gaze-based Looking While Listening measure, showed high reliability across English and Spanish versions (median = 0.94, range = 0.77-0.99). Executive Function/Memory and Math measures showed moderate to high reliability (median = 0.74, range = 0.65-0.87). Due to a testing error during data collection, analysis results for the MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests (CAT) Production measure in Spanish are not reported in this technical manual, here and elsewhere.

4.1.2.1 Empirical Reliability Overall and by Age

Table 4.1

Language Measures (English)

Measure	Age range (months)	n	CSS R ₁₁	SEM (CSS)
Looking While Listening	6-24	379	0.60	5.44
Picture Vocabulary	25+	684	0.88	3.72
MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests (CAT) Production	9-30	1,125	0.99	2.43
MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests (CAT) Comprehension	7-18	649	0.98	2.21
Mullen Receptive	1-48	1,970	0.92	5.63
Mullen Expressive	1-48	1,964	0.77	4.94

Table 4.2

Language Measures (Spanish)

Measure	Age range (months)	n	CSS R ₁₁	SEM (CSS)
Picture Vocabulary	25+	196	0.86	3.72
MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests (CAT) Comprehension	7-18	76	0.96	2.10
Mullen Receptive	1-48	452	0.91	6.04
Mullen Expressive	1-48	448	0.78	4.86

Table 4.3

Executive Function/Memory Measures (English and Spanish)

Measure	Age range (months)	n	CSS R ₁₁	SEM (CSS)
Mullen Visual Reception	1-48	2351	0.65	5.85
Visual Delayed Response	22-48	972	0.77	6.61
Delayed Memory (Learning)	22-48	932	0.79	6.66
Delayed Memory (Test)	22-48	658	0.78	6.45

Table 4.4*Math Measures (English and Spanish)*

Measure	Age range (months)	n	CSS R ₁₁	SEM (CSS)
Who Has More	25-48	819	0.87	5.30
Subitizing	25-48	809	0.67	10.70
Verbal Arithmetic	37-48	395	0.70	9.55
Counting	25-48	741	0.67	5.85

4.1.2.2 Test-Retest Reliability

Test-retest reliability describes how well a measure score remains consistent in measuring the individual's performance across multiple administrations (Revelle & Condon, 2019). A subsample of $n = 220$ total NIH Baby Toolbox norming children and their caregivers were recruited to complete the battery after 1-14 days following the first session ($n = 196$ tested, median = 5 days). Table 4.5 shows the Pearson correlations, absolute agreement intraclass correlation coefficients (ICCs), mean-score differences, and their associated standardized mean difference (SMD) for the scores between the first and second administrations of the NIH Baby Toolbox battery. Note the differences in sample size by measure completion due to differences in appropriate age bands for each measure. SMD was calculated as the absolute value score mean difference divided by the baseline standard deviation across all participants who completed the measure during the first session. (Test-retest statistics were not reported for sample sizes of less than $n = 20$.)

Overall, measures in the Language (English), Executive Function/Memory, and Math domains had moderate to high test-retest reliabilities (Pearson $r = 0.51$ - 0.89) except for the Spanish version of CDI-CAT Production within the Language subdomain, likely due to a much smaller sample size. A larger SMD > 0.20 was also evident in several measure scores, including the MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests (CAT) Comprehension (English), Mullen Receptive (English), Mullen Receptive and Expressive (Spanish), and Subitizing measures.

Table 4.5*Test-Retest Reliability for Cognition Domain Measures*

Domain	Measure	n	Pearson r	95% CI for r	ICC	95% CI for ICC	SMD
Language (English)	Looking While Listening (English)	21	0.74	[.45, .89]	0.74	[.47, .89]	.11
	Picture Vocabulary (English)	36	0.83	[.70, .91]	0.83	[.69, .91]	<.01

	MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests (CAT) Production (English)	80	0.82	[.74, .88]	0.82	[.74, .88]	0.17
	MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests (CAT) Comprehension (English)	57	0.78	[.65, .86]	0.73	[.58, .83]	0.38
	Mullen Receptive (English)	156	0.89	.85, .92]	0.89	[.85, .92]	0.26
	Mullen Expressive (English)	158	0.84	[.79, .88]	0.84	[.78, .88]	0.17
Language (Spanish)	Looking While Listening (Spanish)	0	-	-	-	-	-
	Picture Vocabulary (Spanish)	5	-		-		-
	MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests (CAT) Comprehension (Spanish)	4	-		-		-
	Mullen Receptive (Spanish)	27	0.87	[.72, .94]	0.85	[.71, .93]	0.59
	Mullen Expressive (Spanish)	24	0.88	[.73, .95]	0.88	[.74, .95]	0.64
Executive Function/ Cognition	Mullen Visual Reception	177	0.70	[.62, .77]	0.69	[.61, .76]	0.17
	Visual Delayed Response	40	0.71	[.51, .84]	0.71	[.52, .84]	0.07
	Delayed Memory (Learning)	46	0.62	[.25, .85]	0.62	[.26, .84]	0.02
	Delayed Memory (Test)	18	-	-	-	-	-
Math	Who Has More	39	0.60	[.35, .77]	0.60	[.36, .77]	0.18
	Subitizing	34	0.66	[.41, .82]	0.66	[.49, .82]	0.28
	Verbal Arithmetic	20	0.51	[.08, .78]	0.52	[.12, .78]	0.19
	Counting	41	0.69	[.48, .82]	0.69	[.49, .82]	0.04

4.1.3 Motor Domain Reliability

Table 4.6 shows the empirical reliability coefficients for the Motor domain measures that were scored using an IRT metric and had a CSS. Of note, the Reach to Eat (Block) version of the measure was not intended to be administered to children over 10 months old, though results in Table 4.6 include the full sample of children the measure was administered to. Table 4.7 shows the Pearson correlations, absolute agreement ICCs, and mean-score differences and associated SMD statistics for the Motor domain scores from the first and second administrations of each measure. Overall, empirical reliability for Motor domain measures was high (0.74-0.91). Test-retest reliability for Motor domain measures was also high (0.84-0.97) with the exception of Manual Planning, which had a

smaller retest sample size. SMD was low <0.10 for all measures, indicating little to no practice effects.

4.1.3.1 Empirical Reliability Coefficients Overall and by Age

Table 4.6

Empirical Reliability Coefficients for IRT-Based CSS for Motor Domain Measures

Measure	Score	Age range (months)	n	CSS R_{11}	SEM (CSS)
Reach to Eat (Cheerio)	Fine Motor	1-48	1,262	0.91	5.26
Reach to Eat (Manual Dexterity, Cheerio)	Manual Dexterity	1-48	1,266	0.89	5.68
Reach to Eat (Manual Planning, Cheerio)	Manual Planning	1-48	311	0.74	7.88

4.1.3.2 Test-Retest Reliability

Table 4.7

Test-Retest Reliability for Motor Domain Measures

Measure	Score	n	Pearson r	95% CI for r	ICC	95% CI for ICC	SMD
Get Up and Go	Gross Motor	159	0.90	[.86, .92]	0.89	[.86, .92]	0.06
	Locomotion	166	0.97	[.96, .98]	0.97	[.95, .97]	0.05
	Postural Transition	163	0.84	[.79, .88]	0.84	[.79, .88]	0.09
Reach to Eat (Cheerio)	Fine Motor	56	0.90	[.83, .94]	0.89	[.83, .94]	0.06
Reach to Eat (Fine Motor)	Manual Dexterity	56	0.88	[.80, .93]	0.87	[.80, .93]	0.05
Reach to Eat (Manual Dexterity, Cheerio)	Manual Planning	11	-	-	-	-	-
Reach to Eat (Block)	Manual Dexterity	54	0.84	[.73, .90]	0.83	[.73, .90]	0.05

4.1.4 Social-Emotional Functioning Domain Reliability

Table 4.8 shows the empirical reliability coefficients for the Social-Emotional Functioning domain measures that were scored using an IRT metric and had a CSS. Table 4.9 shows the Pearson correlations, absolute agreement ICCs, and mean-score

differences and associated SMD statistics for the Social-Emotional Functioning domain scores from the first and second administrations of each measure. Overall, empirical reliability for Social-Emotional Functioning domain measures was high (median = 0.82, range = 0.78-0.90) except for the PROMIS Child-Caregiver Interactions, which had a moderate reliability of 0.65. Test-retest reliability for Social-Emotional Functioning domain measures was moderate to high (median = 0.75, range = 0.50-0.86). Larger effect sizes of SMD > 0.20 were evident in several measure scores, including SOM–Younger (9-23 months old) Play, SOM–Older (24-42 months old), and PROMIS Caregiver-Child Interactions.

4.1.4.1 Empirical Reliability Coefficients Overall and by Age

Table 4.8

Empirical Reliability Coefficients for IRT-Based CSS for Social-Emotional Functioning Domain Measures

Measure	Score	Age range tested (months)	n	CSS R ₁₁	SEM (CSS)
Infant Behavior Questionnaire - Revised - Very Short Form (IBQ-R-VSF)	Effortful Control	3-12	695	0.84	4.01
	Negative Affect			0.80	4.42
	Surgency			0.90	3.12
	Stranger Danger			0.79	4.37
Early Childhood Behavior Questionnaire - Very Short Form (ECBQ-VSF)	Effortful Control	13-36	1,066	0.82	4.18
	Negative Affect			0.82	4.20
	Surgency			0.80	4.56
Children's Behavior Questionnaire - Very Short Form (CBQ-VSF)	Effortful Control	37-48	337	0.84	3.85
	Negative Affect			0.86	3.51
	Surgency			0.83	4.09
Caregiver Checklist	-	6-23	941	0.80	4.29
	-	24-48	860	0.72	4.84
Social Observation	Play	9-23	760	0.85	3.69
	Shares Attention			0.79	4.64
	Shares Enjoyment			0.79	4.62
	Shares Interest			0.83	4.06
	-	24-48	803	0.78	4.66
PROMIS EC Parent-Report Bank v1.0 - Anger/Irritability	-	13-48	1,685	0.89	3.35
PROMIS EC Parent-Report Bank v1.0 - Anxiety	-	13-48	1,686	0.81	3.63
PROMIS EC Parent-Report SF v1.0 - Soc Rel - Child-Caregiver Interactions	-	13-48	1,688	0.65	5.01
PROMIS EC Parent-Report Bank v1.0 - Depressive Symptoms	-	13-48	1,686	0.78	3.64

PROMIS EC Parent-Report SF v1.0 - Soc Rel - Peer Relationships 4a	-	13-48	1,687	0.81	3.83
PROMIS EC Parent-Report Bank v1.0 - Positive Affect	-	13-48	1,686	0.86	3.73
PROMIS EC Parent-Report Scale v1.0 - Self-Reg - Flexibility 5a	-	13-48	1,683	0.88	3.33
PROMIS EC Parent-Report Scale v1.0 - Self-Reg - Frustration Tolerance 6a	-	13-48	1,685	0.89	3.31

Note. The PROMIS utilized the T-Score towards calculation of empirical reliability instead of CSS. *

4.1.4.2 Test-Retest Reliability

Table 4.9

Test-Retest Reliability for Social-Emotional Functioning Domain Measures

Measure	Score	Age range (months)	n	Pearson r	95% CI for r	ICC	95% CI for ICC	SMD
Infant Behavior Questionnaire - Revised - Very Short Form (IBQ-R-VSF)	Effortful Control	3-12	74	0.81	[.71, .87]	0.80	[.70, .87]	0.05
	Negative Affect			0.75	[.62, .83]	0.74	[.62, .83]	0.08
	Surgency			0.78	[.67, .86]	0.78	[.68, .86]	0.05
	Stranger Danger			0.77	[.66, .85]	0.77	[.66, .85]	0.06
Early Childhood Behavior Questionnaire - Very Short Form (ECBQ-VSF)	Effortful Control	13-36	64	0.83	[.74, .89]	0.81	[.71, .88]	0.01
	Negative Affect			0.83	[.73, .89]	0.82	[.72, .88]	0.18
	Surgency			0.81	[.70, .88]	0.79	[.68, .87]	0.12
Children's Behavior Questionnaire - Very Short Form (CBQ-VSF)	Effortful Control	37-48	18	-	-	-	-	-
	Negative Affect			-	-	-	-	-
	Surgency			-	-	-	-	-
Caregiver Checklist	-	6-23	82	0.85	[.78, .90]	0.85	[.77, .90]	<0.01
	-	24-48	41	0.75	[.57, .86]	0.75	[.58, .86]	0.06
Social Observation	Play	9-23	59	0.50	[.27, .67]	0.48	[.26, .66]	0.23
	Shares Attention			0.50	[.28, .67]	0.50	[.50, .28]	0.07
	Shares Enjoyment			0.56	[.35, .71]	0.56	[.35, .71]	0.11
	Shares Interest			0.50	[.28, .67]	0.49	[.27, .66]	0.20
	-	24-48	37	0.81	[.66, .90]	0.79	[.63, .88]	0.26
PROMIS EC Parent-Report Bank v1.0 - Anger/Irritability	-	13-48	92	0.79	[.70, .86]	0.79	[.70, .86]	0.06
PROMIS EC Parent-Report Bank v1.0 - Anxiety	-	13-48	93	0.59	[.44, .71]	0.58	[.42, .70]	0.18

PROMIS EC Parent-Report SF v1.0 - Soc Rel - Child-Caregiver Interactions	-	13-48	93	0.53	[.36, .66]	0.50	[.33, .63]	0.28
PROMIS EC Parent-Report Bank v1.0 - Depressive Symptoms	-	13-48	93	0.54	[.38, .67]	0.54	[.38, .67]	0.08
PROMIS EC Parent-Report SF v1.0 - Soc Rel - Peer Relationships 4a	-	13-48	93	0.63	[.49, .74]	0.62	[.48, .73]	0.12
PROMIS EC Parent-Report Bank v1.0 - Positive Affect	-	13-48	93	0.58	[.42, .70]	0.58	[.43, .70]	0.03
PROMIS EC Parent-Report Scale v1.0 - Self-Reg - Flexibility 5a	-	13-48	92	0.73	[.62, .82]	0.74	[.63, .82]	0.01
PROMIS EC Parent-Report Scale v1.0 - Self-Reg - Frustration Tolerance 6a	-	13-48	93	0.71	[.59, .80]	0.70	[.58, .79]	0.09

Note. As stated above, test-retest reliability was not computed for sample sizes ≤ 20 .

4.1.5 Composites

Composites utilized the Mosier composite reliability following Equation 4.3 (Mosier, 1943):

$$\text{Equation 4.3: } p_{xx} = \frac{\bar{p}_{x_i x_i} k + k(k-1) \bar{p}_{x_i x_j}}{k + k(k-1) \bar{p}_{x_i x_j}}$$

Where $\bar{p}_{x_i x_i}$ is the mean reliability of variables in the composites, $\bar{p}_{x_i x_j}$ is the mean intercorrelation among variables in the composite, and k is the number of variables in the composite. See Table 4.10 for reliability coefficients for composite scores.

4.1.5.1 Empirical Reliability Coefficients Overall and by Age

Table 4.10

Reliability Coefficients for Composite Scores

Composite	Age range (months)	Composite reliability (Mosier)
Executive Function/ Memory	22-42	0.87
Math	25-42	0.86
Mullen Language	1-42	0.92
Cognition	25-42	0.94

Motor	1-42	0.88
Negative Affect	13-42	0.94
Self-Regulation	13-42	0.92
Social Communication	9-23	0.91

4.1.5.2 Test-Retest Reliability

Table 4.11

Test-Retest Reliability for Composite Scores

Composite	n	Pearson r	95% CI for r	ICC	95% CI for ICC	SMD
Executive Function/ Memory	17	0.63	[.21, .85]	0.62	[.23, .84]	0.05
Math	30	0.84	[.69, .92]	0.83	[.67, .91]	0.19
Mullen Language	175	0.92	[.89, .94]	0.92	[.89, .94]	0.03
Cognition	12	0.94	[.78, .98]	0.92	[.77, .98]	0.17
Motor	99	0.96	[.94, .97]	0.96	[.94, .97]	0.03
Negative Affect	89	0.79	[.70, .86]	0.79	[.69, .85]	0.12
Self-Regulation	91	0.78	[.68, .85]	0.78	[.68, .85]	0.07
Social Communication	59	0.61	[.41, .75]	0.59	[.39, .96]	0.24

4.1.5.3 Composite Standard Error of Measurement

Because the individuals in the subsample who participated in the test-retest study had a different (restricted) variance of scores than the overall population, there is a risk of bias in the calculations of the composite standard error of measurement (SEM). Therefore, the reliability was corrected following Equation 4.3 (Revelle, 2015; Stauffer & Mendoza, 2001; Thorndike, 1949):

$$\text{Equation 4.3: } R = \frac{\left(\frac{rS}{s}\right)}{\sqrt{\left(1-r^2+r^2\left(\frac{S^2}{s^2}\right)\right)}}$$

Where R is the unrestricted ICC, r is the restricted ICC, S is the unrestricted standard deviation across all participants at baseline, and s is the restricted standard deviation using baseline data from participants with a retest case only.

Composite-level SEMs then follows Equation 4.4:

$$\text{Equation 4.4: } SEM_{\text{composite}} = \frac{SD}{\sqrt{(1-ICC)}}$$

Calculations of the composite SEM utilizing the unrestricted ICC and restricted ICC are provided in Table 4.12 for comparison. The final composite-level SEMs implemented in the NIH Baby Toolbox utilizes SEM using SD_{all} and $ICC_{restricted}$.

Table 4.12

Comparisons of Composite-Level SEM Calculated with Restricted ICC and Unrestricted ICC for Each Composite

Composite	ICC_1	$ICC_{restricted}$	SEM using SD_{all} and ICC_1	SEM using SD_{all} and $ICC_{restricted}$
Cognition	0.92	0.96	2.19	1.67
Executive Function/Memory	0.62	0.55	7.11	7.72
Mullen Language	0.92	0.92	4.04	4.12
Math	0.83	0.83	5.41	5.32
Motor	0.96	0.95	3.17	3.61
Negative Affect	0.79	0.77	6.18	6.42
Self-Regulation	0.78	0.75	6.86	7.17
Social Communication	0.59	0.67	7.83	7.01

4.2 Validity

4.2.1 Developmental Change in Measures and Composites by Age

As many of these domain measures are associated with development, it was expected that there would be improvements in performance in many, but not all, participants as age increases. Tables 4.13-4.15 show the Spearman rho (ρ) correlations between Cognition, Motor, and Social-Emotional Functioning domain measure scores and age in months. Table 4.16 shows the Spearman ρ correlations between the composite scores and chronological age in months rounded to the nearest month. Additionally, bootstrapped 95% confidence intervals (using 1,000 bootstraps) around Spearman ρ are provided. As expected, most Cognition and Motor domain measures are positively correlated with age. Social-Emotional Functioning domain measures do not, nor were they expected to, correlate uniformly with age.

Table 4.13

Spearman rho (ρ) Correlations Between Cognition Measure Scores and Chronological Age

Domain	Measure	n	Spearman ρ	95% bootstrapped CI
Language (English)	Looking While Listening (English)	379	.39	[.30, .47]
	Picture Vocabulary (English)	684	.53	[.48, .59]
	MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests (CAT) Production (English)	1,125	.80	[.77, .82]
	MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests (CAT) Comprehension (English)	649	.66	[.61, .70]
	Mullen Receptive (English)	1,970	.87	[.85, .88]
	Mullen Expressive (English)	1,964	.87	[.85, .88]
Language (Spanish)	Looking While Listening (Spanish)	42	.14	[-.17, .43]
	Picture Vocabulary (Spanish)	196	.40	[.26, .51]
	MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests (CAT) Production (Spanish)	154	.63	[.48, .73]
	MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests (CAT) Comprehension (Spanish)	76	.63	[.48, .75]
	Mullen Receptive (Spanish)	452	.85	[.82, .88]
	Mullen Expressive (Spanish)	448	.85	[.83, .88]
Executive Function/ Memory	Mullen Visual Reception	2,351	.63	[.59, .66]
	Familiarization (Mean Looking) *	319	-.28	[-.38, -.18]
	Familiarization (Sum Looking) *	319	-.28	[-.38, -.16]
	Familiarization (Peak Looking) *	319	-.22	[-.32, -.12]
	Visual Delayed Response	972	.46	[.41, .51]
	Delayed Memory (Learning)	932	.42	[.36, .47]
	Delayed Memory (Test)	658	.22	[.14, .29]
Math	Who Has More	819	.45	[.39, .51]
	Subitizing	809	.47	[.41, .52]
	Verbal Arithmetic	395	.25	[.13, .34]
	Counting	741	.51	[.46, .56]

Note. For measures with an asterisk (*), negative correlation coefficients are expected. Scores for Familiarization also span Both Rounds.

Table 4.14

Spearman rho (ρ) Correlations Between Motor Measure Scores and Chronological Age in Months

Measure	Score	n	Spearman ρ	95% bootstrapped CI
Get Up and Go	Gross Motor	2,145	.85	[.84, .86]
	Locomotion	2,164	.84	[.82, .85]
	Postural Transition	2,173	.84	[.82, .86]
Reach to Eat (Cheerio)	Fine Motor	1,268	.64	[.60, .67]
	Manual Dexterity	1,266	.61	[.57, .65]
	Manual Planning	311	.32	[.20, .42]
Reach to Eat (Block)	Manual Dexterity	466	.79	[.74, .83]

Table 4.15

Spearman rho (ρ) Correlations Between Social-Emotional Functioning Measure Scores and Chronological Age in Months

Measure	Score	Age range (months)	n	Spearman ρ	95% bootstrapped CI
Infant Behavior Questionnaire - Revised - Very Short Form (IBQ-R-VSF)	Effortful Control	3-12	695	.04	[-.05, .11]
	Negative Affect			.06	[-.02, .13]
	Surgency			.60	[.54, .66]
	Stranger Danger			.48	[.42, .53]
Early Childhood Behavior Questionnaire - Very Short Form (ECBQ-VSF)	Effortful Control	13-36	1,066	.22	[.16, .27]
	Negative Affect			.20	[.14, .26]
	Surgency			.16	[.09, .21]
Children's Behavior Questionnaire - Very Short Form (CBQ-VSF)	Effortful Control	37-48	337	.07	[-.05, .17]
	Negative Affect			.01	[-.09, .12]
	Surgency			.08	[-.03, .18]
Caregiver Checklist	-	6-23	941	.66	[.62, .70]
	-	24-48	860	.04	[-.04, .09]
Social Observation	Play	9-23	760	.55	[.49, .60]
	Shares Attention			.004	[-.06, .08]
	Shares Enjoyment			.12	[.05, .19]
	Shares Interest			.40	[.33, .46]
	-	24-48	803	.49	[.43, .55]
PROMIS EC Parent-Report Bank v1.0 - Anger/Irritability	-	13-48	1,685	.13	[.08, .18]

PROMIS EC Parent-Report Bank v1.0 - Anxiety	-	13-48	1,686	.07	[.02, .11]
PROMIS EC Parent-Report SF v1.0 - Soc Rel - Child-Caregiver Interactions	-	13-48	1,688	.13	[.09, .17]
PROMIS EC Parent-Report Bank v1.0 - Depressive Symptoms	-	13-48	1,686	.08	[.03, .13]
PROMIS EC Parent-Report SF v1.0 - Soc Rel - Peer Relationships 4a	-	13-48	1,687	.13	[.08, .17]
PROMIS EC Parent-Report Bank v1.0 - Positive Affect	-	13-48	1,686	.02	[-.03, .07]
PROMIS EC Parent-Report Scale v1.0 - Self-Reg - Flexibility 5a	-	13-48	1,683	.004	[-.04, .05]
PROMIS EC Parent-Report Scale v1.0 - Self-Reg - Frustration Tolerance 6a	-	13-48	1,685	.14	[.09, .19]

Table 4.16

Spearman rho (ρ) Correlations Between Composite Scores and Chronological Age in Months

Composite	n	Spearman ρ	95% bootstrapped CI
Executive Function/ Memory	601	.43	[.36, .50]
Math	605	.59	[.53, .64]
Mullen Language	2,360	.90	[.89, .90]
Cognition	383	.63	[.57, .69]
Motor	1,539	.87	[.85, .88]
Negative Affect	1,357	.02	[-.03, .07]
Self-Regulation	1,382	.02	[-.04, .07]
Social Communication	743	.60	[.55, .64]

4.2.2 Relationship of the NIH Baby Toolbox to other Measures

4.2.2.1 Bayley-4 Scales of Infant and Toddler Development

A subset of infants (n = 120 recruited) returned 1-14 days after the first session to complete the Bayley-4 Scales of Infant and Toddler Development assessment (Bayley & Aylward, 2019), which is a norm-referenced developmental assessment that has been

used for several decades in clinical, educational, and research settings. The Bayley-4 assessment evaluates Cognition, Expressive Language, Receptive Language, Gross Motor, and Fine Motor scales, which contain polytomous items with response to categories ranging from 0 to 2. The Bayley-4 assessment also includes caregiver-report measures for Social-Emotional Functioning scales and Adaptive Behavior scales, assessing Receptive, Expressive, Personal, Interpersonal Relationships, and Play and Leisure subscales. Growth scale values (GSVs) provide an estimate of the level of underlying ability that is most likely to have produced that raw score. Pearson correlations between the NIH Baby Toolbox measure scores and Bayley-4 GSVs from the conceptually closest convergent domain are available in Tables 4.17-4.20. Norm-referenced scaled score norm values for the Cognitive, Expressive Language, Receptive Language, Gross Motor, and Fine Motor domains, as well as Social-Emotional Functioning and Adaptive Behavior scales, were obtained to evaluate convergent validity against the NIH Baby Toolbox measure-level scores. Pearson correlations between the NIH Baby Toolbox measure and composite norm scores and Bayley-4 assessment conceptually closest convergent scaled score norm values are available in Table 4.21.

Table 4.17

Pearson Correlations Between NIH Baby Toolbox Cognition Measure Scores and Bayley-4 GSVs

Domain	Measure	Bayley-4 domain	Overlapping n	Pearson r
Language (English)	Looking While Listening (English)	Receptive Language	18	0.49
	Picture Vocabulary (English)	Receptive Language	40	0.57
	MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests (CAT) Production (English)	Expressive Language	53	0.77
	MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests (CAT) Comprehension (English)	Receptive Language	34	0.80
	Mullen Receptive (English)	Receptive Language	109	0.90
	Mullen Expressive (English)	Expressive Language	109	0.91
Executive Function/ Memory	Mullen Visual Reception	Cognition	112	0.75
	Familiarization (Mean Looking Both Rounds)	Cognition	18	-0.51*
	Familiarization (Sum Looking Both Rounds)	Cognition	18	-0.51*
	Familiarization (Peak Looking Both Rounds)	Cognition	18	-0.58*
	Visual Delayed Response	Cognition	50	0.46

Math	Delayed Memory (Learning)	Cognition	41	0.39
	Delayed Memory (Test)	Cognition	35	0.21
	Who Has More	Cognition	41	0.26
	Subitizing	Cognition	41	0.64
	Verbal Arithmetic	Cognition	20	0.46
	Counting	Cognition	44	0.68

Note. For measures with an asterisk (*), negative correlation coefficients are expected.

Table 4.18

Pearson Correlations Between NIH Baby Toolbox Motor Measure Scores and Bayley-4 GSVs

Measure	Score	Bayley-4 domain	Overlapping n	Pearson r
Get Up and Go	Gross Motor	Gross Motor	106	0.89
	Locomotion	Gross Motor	107	0.94
	Postural Transition	Gross Motor	106	0.89
Reach to Eat (Cheerio)	Fine Motor	Fine Motor	61	0.68
	Manual Dexterity	Fine Motor	61	0.70
	Manual Planning	Fine Motor	17	0.31
Reach to Eat (Block)	Manual Dexterity	Fine Motor	25	0.51

Table 4.19

Pearson Correlations Between NIH Baby Toolbox Social-Emotional Functioning Measure Scores and Bayley-4 GSVs

Measure	Score	Age range (months)	Bayley-4 domain	Overlapping n	Pearson r
Infant Behavior Questionnaire - Revised - Very Short Form (IBQ-R-VSF)	Effortful Control	3-12	Cognition	42	-0.002
	Negative Affect		Interpersonal	39	-0.24
			Personal	40	-0.31
			Play and Leisure	40	-0.22
	Surgency		Interpersonal	39	0.47
			Personal	40	0.48
			Play and Leisure	40	0.55
			Receptive	40	0.67
	Stranger Danger		Interpersonal	39	0.60
Early Childhood Behavior Questionnaire - Very Short	Effortful Control	13-36	Cognition	53	0.21
	Negative Affect		Interpersonal	52	-0.39
			Personal	52	-0.21

Form (ECBQ-VSF)			Play and Leisure	52	-0.34
	Surgency		Interpersonal	52	0.39
			Personal	52	0.31
			Play and Leisure	52	0.25
Children's Behavior Questionnaire - Very Short Form (CBQ-VSF)	Effortful Control	37-48	Cognition	20	0.26
	Negative Affect		Interpersonal	20	-0.35
			Personal	20	-0.43
			Play and Leisure	20	-0.33
	Surgency		Interpersonal	20	0.52
			Personal	20	0.59
			Play and Leisure	20	0.42
Caregiver Checklist	-	6-23	Expressive	44	0.65
			Receptive	45	0.61
			Interpersonal	45	0.53
			Play and Leisure	45	0.59
	-	24-48	Expressive	47	0.55
			Receptive	47	0.65
			Interpersonal	47	0.58
			Play and Leisure	47	0.62
Social Observation	Play	9-23	Expressive	28	0.67
			Receptive	28	0.54
			Interpersonal	28	0.47
			Play and Leisure	28	0.55
	Shares Attention		Expressive	28	0.004
			Receptive	28	0.01
			Interpersonal	28	-0.09
	Shares Enjoyment		Expressive	28	0.51
			Receptive	28	0.47
			Interpersonal	28	0.50
	Shares Interest		Expressive	28	0.68
			Receptive	28	0.68
			Interpersonal	28	0.52
	-	24-48	Expressive	39	0.67
			Receptive	39	0.53
			Interpersonal	39	0.49
			Play and Leisure	39	0.47
PROMIS EC Parent-Report Bank v1.0 - Anger/Irritability	-	13-48	Interpersonal	73	-0.24
			Personal	73	-0.19
			Play and Leisure	73	-0.26
	-	13-48	Interpersonal	73	-0.07

PROMIS EC Parent-Report Bank v1.0 - Anxiety			Personal	73	0.003
			Play and Leisure	73	0.03
PROMIS EC Parent-Report SF v1.0 - Soc Rel - Child-Caregiver Interactions	-	13-48	Expressive	73	0.16
			Receptive	73	0.25
			Interpersonal	73	0.32
			Play and Leisure	73	0.37
PROMIS EC Parent-Report Bank v1.0 - Depressive Symptoms	-	13-48	Interpersonal	73	-0.23
			Personal	73	-0.13
			Play and Leisure	73	-0.14
PROMIS EC Parent-Report SF v1.0 - Soc Rel - Peer Relationships 4a	-	13-48	Expressive	72	0.23
			Receptive	72	0.24
			Interpersonal	72	0.43
			Play and Leisure	72	0.33
PROMIS EC Parent-Report Bank v1.0 - Positive Affect	-	13-48	Interpersonal	73	0.32
			Personal	73	0.24
			Play and Leisure	73	0.21
PROMIS EC Parent-Report Scale v1.0 - Self-Reg - Flexibility 5a	-	13-48	Interpersonal	73	0.32
			Personal	73	0.31
			Play and Leisure	73	0.31
			Cognition	74	0.21
PROMIS EC Parent-Report Scale v1.0 - Self-Reg - Frustration Tolerance 6a	-	13-48	Interpersonal	73	0.36
			Personal	73	0.30
			Play and Leisure	73	0.39
			Cognition	74	0.16

Note. Except for Cognition, all Bayley-4 domains here are from the Adaptive Behavior Scale (ADBE). For measures with an asterisk (*), negative correlation coefficients are expected.

Table 4.20

Pearson Correlations Between NIH Baby Toolbox Composite Scores and Bayley-4 GSVs

Composite or measure	Bayley-4 domain	Overlapping n	Pearson r
Executive Function/Memory	Cognition	32	0.40
Math	Cognition	40	0.73
Mullen Language	Receptive Language	113	0.93
	Expressive Language	113	0.93
Cognition	Cognition	29	0.76
Motor	Fine Motor	79	0.88
	Gross Motor	79	0.92

Negative Affect	Interpersonal	69	-0.35
	Personal	70	-0.24
	Play and Leisure	70	-0.30
	Cognition	71	-0.07
Self-Regulation	Interpersonal	71	0.43
	Personal	72	0.36
	Play and Leisure	72	0.47
	Cognition	73	0.17
Social Communication	ADBE Receptive	26	0.58
	ADBE Expressive	26	0.69
	Interpersonal	26	0.52
	Play and Leisure	26	0.54

Table 4.21

Pearson Correlations Between NIH Baby Toolbox Measure and Composite Normed Standard Scores and Bayley-4 Normed Scaled Scores

Composite or measure	Bayley-4 domain	Overlapping n	Pearson r
Picture Vocabulary (English)	Receptive Language	21	0.44
Mullen Expressive (English)	Expressive Language	98	0.50
Mullen Receptive (English)	Receptive Language	99	0.53
Fine Motor	Fine Motor	52	0.04
Executive Function/ Memory	Cognition	24	0.39
Math	Cognition	29	0.58
Mullen Language	Receptive Language	103	0.52
	Expressive Language	103	0.54
Cognition	Cognition	21	0.65
Motor	Fine Motor	71	-0.07
	Gross Motor	71	0.38
Negative Affect	Cognition	61	-0.03
	Social-Emotional	60	-0.47
Self-Regulation	Cognition	63	0.08
	Social-Emotional	62	0.42
Social Communication	Receptive Language	26	0.53
	Expressive Language	26	0.31
	Social-Emotional	25	0.55

4.2.2.2 Ages and Stages Questionnaire (ASQ)

Caregivers of the child participants (n = 2,515 recruited) completed the Ages and Stages Questionnaire 3rd edition (ASQ-3) (Squires & Bricker, 2009), a criterion-referenced caregiver-report form that assesses Communication, Gross Motor, Fine Motor, Personal Social, and Problem-Solving domains. The ASQ-3 has been validated in medical literature and has a sensitivity of about 75% to detect delayed development (Squires et al., 1997). Each domain contains six age-specific items to which caregivers can respond yes, sometimes, or no (scored as 10, 5, or 0 points, respectively).

Caregivers are also given the opportunity to provide more information about the child in a supplementary section that is not scored. The ASQ-3 was scored, and cutoff categories were assigned as per its standardization (Squires & Bricker, 2009). Children are assigned to one of three categories, defined as on-schedule, monitoring zone, or referral. The monitoring zone, as per standardization, was defined as scores that were greater than 1 and less than 2 standard deviations below the mean. Children scoring below this range are classified as requiring further evaluation and referral, while children scoring above are classified as on-schedule.

One-way, unbalanced analyses of covariance (ANCOVAs) were conducted between the NIH Baby Toolbox scores and the closest convergent ASQ-3 domain score categories. Chronological age was mean centered and used as a covariate in the ANCOVA model. Results are available in Tables 4.22-4.25.

Table 4.22

One-Way ANCOVA Results Between NIH Baby Toolbox Cognition Measure Scores and ASQ-3 Domain Categories

Domain	Measure	ASQ-3 domain	Overlapping n	F	p-value	Partial η^2
Language (English)	Looking While Listening (English)	Communication	372	4.42	0.01	0.02
	Picture Vocabulary (English)	Communication	679	25.35	<.001	0.05
	MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests (CAT) Production (English)	Communication	1,109	70.97	<.001	0.07
	MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests	Communication	636	23.86	<.001	0.05

	(CAT) Comprehension (English)					
	Mullen Receptive (English)	Communication	1,940	104.30	<.001	0.03
	Mullen Expressive (English)	Communication	1,936	154.70	<.001	0.04
Language (Spanish)	Looking While Listening (Spanish)	Communication	42	3.31	0.05	0.15
	Picture Vocabulary (Spanish)	Communication	189	3.13	0.05	0.03
	MacArthur-Bates Communicative Development Inventories (CDI) Computerized Adaptive Tests (CAT) Comprehension (Spanish)	Communication	75	0.88	0.42	0.02
	Mullen Receptive (Spanish)	Communication	442	3.88	0.02	<.01
	Mullen Expressive (Spanish)	Communication	436	11.19	<.001	0.02
Executive Function/ Cognition	Mullen Visual Reception	Problem-Solving	2,298	58.08	<.001	0.03
	Familiarization (Mean Looking Both Rounds)	Problem-Solving	311	3.93	0.02	0.02
	Familiarization (Sum Looking Both Rounds)	Problem-Solving	311	0.03	0.97	<.01
	Familiarization (Peak Looking Both Rounds)	Problem-Solving	311	3.90	0.02	0.02
	Visual Delayed Response	Problem-Solving	954	11.55	<.001	0.02
	Delayed Memory (Learning)	Problem-Solving	912	12.55	<.001	0.02
	Delayed Memory (Test)	Problem-Solving	646	6.06	0.002	0.02
Math	Who Has More	Problem-Solving	804	8.62	<.001	0.02
	Subitizing	Problem-Solving	793	30.35	<.001	0.06
	Verbal Arithmetic	Problem-Solving	386	7.90	<.001	0.04
	Counting	Problem-Solving	729	15.45	<.001	0.03

Table 4.23

One-Way ANCOVA Results Between NIH Baby Toolbox Motor Measure Scores and ASQ-3 Domain Categories

Measure	Score	ASQ-3 domain	Overlapping n	F	p-value	Partial η^2
Get Up and Go	Gross Motor	Gross Motor	2,098	167	<.001	0.05
	Locomotion	Gross Motor	2,117	163.1	<.001	0.05
	Postural Transition	Gross Motor	2,125	143.3	<.001	0.04
Reach to Eat (Cheerio)	Fine Motor	Fine Motor	1,233	2.63	0.07	0.003
	Manual Dexterity	Fine Motor	1,237	2.56	0.08	0.003
	Manual Planning	Fine Motor	308	0.49	0.62	0.003
Reach to Eat (Block)	Manual Dexterity	Fine Motor	457	2.57	0.08	0.007

Table 4.24

One-Way ANCOVA Results Between NIH Baby Toolbox Social-Emotional Functioning Measure Scores and ASQ-3 Domain Categories

Measure	Score	Age range	ASQ-3 domain	Overlapping n	F	p-value	Partial η^2
Infant Behavior Questionnaire - Revised - Very Short Form (IBQ-R-VSF)	Effortful Control	3-12	Problem-Solving	683	16.39	<.001	0.05
	Negative Affect		Personal-Social	686	0.32	0.73	<0.01
	Surgency		Personal-Social	686	53.15	<.001	0.09
	Stranger Danger		Personal-Social	686	14.44	<.001	0.03
Early Childhood Behavior Questionnaire - Very Short Form (ECBQ-VSF)	Effortful Control	13-36	Problem-Solving	1,039	20.55	<.001	0.04
	Negative Affect		Personal-Social	1,045	3.76	0.02	<0.01
	Surgency		Personal-Social	1,045	14.09	<.001	0.03
Children's Behavior Questionnaire - Very Short Form (CBQ-VSF)	Effortful Control	37-48	Problem-Solving	328	14.09	<.001	0.08
	Negative Affect		Personal-Social	332	4.12	0.02	0.02
	Surgency		Personal-Social	332	5.91	0.003	0.03
Caregiver Checklist	-	6-23	Personal-Social	920	75.93	<.001	0.09
			Communication	924	23.64	<.001	0.03
	-	24-48	Personal-Social	848	88.0	<.001	0.17
			Communication	851	83.21	<.001	0.16

Social Observation	Play	9-23	Personal-Social	741	21.97	<.001	0.04
			Communication	743	2.61	0.07	<0.01
	Shares Attention		Personal-Social	741	2.76	0.06	<0.01
			Communication	743	1.06	0.35	<0.01
	Shares Enjoyment		Personal-Social	741	6.57	0.001	0.02
			Communication	743	2.62	0.07	<0.01
	Shares Interest		Personal-Social	741	14.89	<.001	0.03
				Communication	743	12.38	<.001
	-	24-48	Personal-Social	791	35.28	<.001	0.07
	Communication		794	54.28	<.001	0.10	
PROMIS EC Parent-Report Bank v1.0 - Anger/Irritability	-	13-48	Personal-Social	1,652	2.03	0.13	<0.01
PROMIS EC Parent-Report Bank v1.0 - Anxiety	-	13-48	Personal-Social	1,653	0.09	0.82	<0.01
PROMIS EC Parent-Report SF v1.0 - Soc Rel - Child-Caregiver Interactions	-	13-48	Personal-Social	1,653	47.57	<.001	0.05
PROMIS EC Parent-Report Bank v1.0 - Depressive Symptoms	-	13-48	Personal-Social	1,652	5.70	0.003	<0.01
PROMIS EC Parent-Report SF v1.0 - Soc Rel - Peer Relationships 4a	-	13-48	Personal-Social	1,652	47.57	<.001	0.05
PROMIS EC Parent-Report Bank v1.0 - Positive Affect	-	13-48	Personal-Social	1,653	0.92	0.39	<0.01
PROMIS EC Parent-Report Scale v1.0 - Self-Reg - Flexibility 5a	-	13-48	Problem-Solving	1,640	27.33	<.001	0.03
PROMIS EC Parent-Report Scale v1.0 - Self-Reg - Frustration Tolerance 6a	-	13-48	Problem-Solving	1,642	19.91	<.001	0.02

Table 4.25

One-Way ANCOVA Results Between NIH Baby Toolbox Composite Scores and ASQ-3 Domain Categories

Composite or measure	ASQ-3 domain	Overlapping n	F	p-value	Partial η^2
Executive Function/ Memory	Problem-Solving	591	23.48	<.001	0.06
Math	Problem-Solving	594	30.22	<.001	0.06
Mullen Language	Communication	2,322	172.30	<.001	0.03
Cognition	Problem-Solving	377	36.60	<.001	0.11
Motor	Fine Motor	1,507	37.10	<.001	0.01
	Gross Motor	1,508	131.10	<.001	0.05
Negative Affect	Personal-Social	1,336	4.92	0.007	<0.01
	Communication	1,340	2.58	0.08	<0.01
Self-Regulation	Problem-Solving	1,351	26.59	<.001	0.04
	Personal-Social	1,361	33.0	<.001	0.05
Social Communication	Personal-Social	728	42.27	<.001	0.07
	Communication	730	11.87	<.001	0.02

4.3 Factor Structure for Measure Scores

4.3.1 Measure and Composite Score Intercorrelations

4.3.1.1 Language

Analyzing the relationships among the NIH Baby Toolbox Language domain measures can offer further insights into the test content. Tables 4.26 and 4.27 present the correlations between the NIH Baby Toolbox Language measure scores across the English and Spanish samples.

The correlations among the receptive language measures in English (CDI Comp English, LwL English, PV English, and MR English) and Spanish (CDI Comp Spanish, LwL Spanish, PV Spanish, and MR Spanish) indicate varying degrees of positive relationships across these variables. For the English version of the measures, CDI Comp English was positively correlated with LwL English ($r = 0.22$), PV English ($r = 0.33$), and MR English ($r = 0.33$). LwL English showed a positive correlation with PV English ($r = 0.37$) and MR English ($r = 0.37$), while the strongest correlation was observed between PV English and MR English ($r = 0.52$). CDI Prod English was also positively correlated with ME English ($r = 0.58$), indicating a strong association between these two expressive language measures.

Table 4.26*Intercorrelation Matrix for English Language Measures*

	CDI Prod	CDI Comp	LwL	PV	MR	ME
CDI Prod	1.00					
	(1,125)					
CDI Comp	0.54	1.00				
	(546)	(649)				
LwL	0.28	0.22	1.00			
	(308)	(220)	(379)			
PV	0.26	-	-	1.00		
	(220)	(0)	(0)	(684)		
MR	0.53	0.33	0.37	0.52	1.00	
	(1,048)	(612)	(366)	(660)	(1970)	
ME	0.58	0.30	0.34	0.52	0.83	1.00
	(1,039)	(607)	(363)	(676)	(1924)	(1964)

Note. Pearson correlations (sample size). CDI Prod = MacArthur-Bates Communicative Development Inventories Computerized Adaptive Tests Production, CDI Comp = MacArthur-Bates Communicative Development Inventories Computerized Adaptive Tests Comprehension, LwL = Looking While Listening, PV = Picture Vocabulary, MR = Mullen Receptive, ME = Mullen Expressive.

For the Spanish versions of the measures, CDI Comp Spanish was positively correlated with MR Spanish ($r = 0.29$). PV Spanish was positively correlated with MR Spanish ($r = 0.38$). These findings highlight both shared and distinct relationships between English and Spanish receptive language measures, reflecting underlying similarities and differences in the Language domain constructs being assessed.

Table 4.27*Intercorrelation Matrix for Spanish Language Measures*

	CDI Comp	LwL	PV	MR	ME
CDI Comp	1.00				
	(76)				
LwL	0.04	1.00			
	(23)	(42)			
PV	-	-	1.00		
	(0)	(0)	(196)		
MR	0.29	0.00	0.38	1.00	
	(74)	(42)	(188)	(452)	
ME	0.36	0.19	0.47	0.79	1.00
	(73)	(42)	(194)	(436)	(448)

Note. Pearson correlations (sample size). CDI Comp = MacArthur-Bates Communicative Development Inventories Computerized Adaptive Tests Comprehension, LwL = Looking While Listening, PV = Picture Vocabulary, MR = Mullen Receptive, ME = Mullen Expressive.

4.3.1.2 Executive Function/Memory

Correlation analysis revealed the following relationships between the variables: MTT showed a moderate positive correlation with MTL ($r = 0.33$) and a weaker positive correlation with VDR ($r = 0.29$) and MVR ($r = 0.16$). MTL, in turn, had a stronger positive correlation with VDR ($r = 0.41$) and a weak correlation with MVR ($r = 0.18$). VDR exhibited a moderate positive correlation with MTL ($r = 0.41$) but only a weak positive correlation with MVR ($r = 0.17$). MVR had weak positive correlations with MTT, MTL, and VDR, with values of 0.16, 0.18, and 0.17, respectively.

The sum of looking duration to the center of the screen across both Familiarization rounds (Sum Looking) and the mean of looking duration to the center of the screen across both Familiarization rounds (Mean Looking) were almost perfectly correlated, indicating an almost identical pattern of results. Both variables also showed a strong positive correlation with the peak of looking duration to the center of the screen across both Familiarization rounds (Peak Looking) ($r = 0.95$). As shorter looking durations, rather than longer, are considered faster learning for these scores, all three variables had negative but expected correlations with MVR, with values of -0.19, -0.19, and -0.12, respectively.

Table 4.28

Intercorrelation Matrix for Touch-Based Executive Function/Memory Measures for Children 22 Months Old or Older

	MTT	MTL	VDR	MVR
MTT	1.00 (658)			
MTL	0.33 (613)	1.00 (932)		
VDR	0.29 (636)	0.41 (853)	1.00 (972)	
MVR	0.16 (620)	0.18 (860)	0.17 (893)	1.00 (2,351)

Note. MTT = Memory Task Test portion, MTL = Memory Task Learning portion, VDR = Visual Delayed Response, MVR = Mullen Visual Reception.

Table 4.29

Intercorrelation Matrix for Executive Function/Memory Measures for Children Under 22 Months Old

	Sum Looking Familiarization	Mean Looking Familiarization	Peak Look Familiarization	MVR
Sum Looking Familiarization	1.00			
	(319)			
Mean Looking Familiarization	1.00	1.00		
	(319)	(319)		
Peak Looking Familiarization	0.95	0.95	1.00	
	(319)	(319)	(319)	
MVR	-0.19	-0.19	-0.12	1.00
	(316)	(316)	(316)	(2,351)

Note. MVR = Mullen Visual Reception. Scores for Familiarization span Both Rounds.

4.3.1.3 Math

WHM demonstrated a moderate positive correlation with SUB ($r = 0.39$), VA ($r = 0.30$), and Counting ($r = 0.34$). SUB had a stronger positive correlation with Counting ($r = 0.63$), suggesting a more substantial relationship between these two variables. It also showed moderate positive correlations with VA ($r = 0.32$) and WHM ($r = 0.39$). VA had moderate positive correlations with both Counting ($r = 0.38$) and SUB ($r = 0.32$), but the relationships were weaker compared to those involving Counting.

Table 4.30

Intercorrelation Matrix for Math Measures

	WHM	SUB	VA	Count
WHM	1.00			
	(819)			
SUB	0.39	1.00		
	(746)	(809)		
VA	0.30	0.32	1.00	
	(373)	(371)	(395)	
Count	0.34	0.63	0.38	1.00
	(662)	(644)	(307)	(741)

Note. WHM = Who Has More, SUB = Subitizing, VA = Verbal Arithmetic, Count = Counting

4.3.1.4 Motor

High positive correlations were observed among locomotion, postural transition, and gross motor scores, all ranging between 0.84 and 0.97, indicating that these broader motor abilities are closely linked. Postural control is also strongly correlated with these

scores, especially with gross motor abilities ($r = 0.86$) and locomotion ($r = 0.85$), suggesting that individuals with better control over their posture tend to perform better in both locomotion and gross motor measures.

Dexterity measures from the Reach to Eat (Block) version of the measure, including right hand (BD RH), left hand (BD LH), and average dexterity (BD Avg), show strong correlations with each other, particularly between BD RH and BD Avg ($r = 0.96$), which suggests that hand dexterity is closely related across both hands. From the Reach to Eat (Cheerio) version of the measure, MD (Manual Dexterity) and FM (Fine Motor) skills exhibit particularly strong correlations ($r = 0.99$), as the Fine Motor score uses both Cheerio Dexterity and Manual Planning items.

Table 4.31

Intercorrelation Matrix for Motor Measures

	LOC	PT	GM	BD RH	BD LH	BD Avg	MP	MD	FM
LOC	1.00								
	(2,164)								
PT	0.87	1.00							
	(2,147)	(2,173)							
GM	0.87	0.97	1.00						
	(2,145)	(2,145)	(2145)						
BD RH	0.57	0.53	0.65	1.00					
	(429)	(424)	(422)	(464)					
BD LH	0.56	0.52	0.63	0.82	1.00				
	(424)	(419)	(417)	(455)	(457)				
BD Avg	0.58	0.54	0.66	0.96	0.95	1.00			
	(431)	(426)	(424)	(464)	(457)	(466)			
MP	0.22	0.25	0.27	-	-	-	1.00		
	(283)	(286)	(282)	(0)	(0)	(0)	(311)		
MD	0.58	0.53	0.52	-	-	-	0.37	1.00	
	(1,124)	(1,133)	(1117)	(0)	(0)	(0)	(305)	(1266)	
FM	0.59	0.53	0.53	-	-	-	0.55	0.99	1.00
	(1,122)	(1,131)	(1,115)	(0)	(0)	(0)	(305)	(1,262)	(1,262)

Note. LOC = Get Up and Go Locomotion Score, PT = Get Up and Go Postural Transition Score, GM = Get Up and Go Gross Motor Score, BD RH = Reach to Eat (Block) Dexterity Right Hand Score, BD LH = Reach to Eat (Block) Dexterity Left Hand Score, BD Avg = Reach to Eat (Block) Average of Right and Left Hand Dexterity Scores, MP = Reach to Eat (Cheerio) Manual Planning Score, MD = Reach to Eat (Cheerio) Manual Dexterity Score, FM = Reach to Eat (Cheerio) Fine Motor Score.

4.3.1.5 Social-Emotional Functioning

Negative Affect subdomain measures were positively correlated with each other ($r = 0.33-0.52$; Table 4.32) and negatively correlated with the Psychological Well-Being subdomain measure of positive affect ($r = -0.25 - -0.40$; Table 4.33). Moderate intra-domain correlations for Temperament subdomain measures of effortful control and surgency (IBQ $r = 0.57$, $n = 695$; ECBQ $r = 0.47$, $n = 1,066$; CBQ $r = 0.26$, $n = 337$) and Self-Regulation subdomain measures ($r = 0.60$; $n = 1,683$) were found; to a lesser extent, these subdomains were correlated with each other ($r = 0.32-0.47$; Table 4.34). SOM and CC were positively correlated for each age group (Younger: $r = 0.13-0.52$, $n = 760$; Older: $r = 0.26$, $n = 803$), and the CC was modestly correlated with the two PROMIS EC Social Relationships measures ($r = 0.27-0.44$; Table 4.35).

Table 4.32

Intercorrelations for Negative Affect Scores

	PROMIS Anger	PROMIS Anxiety	PROMIS Depressive Symptoms
PROMIS Anger	1	0.33	0.36
	(1,685)	(1,685)	(1,684)
PROMIS Anxiety	0.33	1	0.52
	(1,685)	(1,686)	(1,685)
PROMIS Depressive Symptoms	0.36	0.52	1
	(1,684)	(1,685)	(1,686)

Table 4.33

Correlations Between Negative and Positive Affect T-Scores

	PROMIS Positive Affect
PROMIS Anger	-0.25
	(1,685)
PROMIS Anxiety	-0.34
	(1,686)
PROMIS Depressive Symptoms	-0.40
	(1,685)

Table 4.34

Correlations Between CBQ (CSS), ECBQ (CSS), and Surgency and PROMIS Self-Regulation Scores (T-Scores)

	PROMIS Self-Regulation Flexibility	PROMIS Self-Regulation Frustration
CBQ Effortful Control	0.39	0.32
	(337)	(337)
CBQ Surgency	0.31	0.19
	(337)	(337)
ECBQ Effortful Control	0.40	0.47
	(1,061)	(1,062)
ECBQ Surgency	0.28	0.17
	(1,061)	(1,062)

Table 4.35

Correlations Between Caregiver Checklists (CSS) and PROMIS Caregiver and Peer T-Scores

	PROMIS Caregiver Interactions	PROMIS Peer Relationships
Caregiver Checklist Older	0.41	0.44
	(859)	(858)
Caregiver Checklist Younger	0.37	0.27
	(546)	(546)

4.3.2 Composite Development

Related measure scores were further factor-analyzed to derive composites. Composite scores were sought that are both theoretically grounded and empirically supported, allowing for a more reliable and nuanced assessment by capturing a broader range of related behaviors and patterns. Combining multiple measures into a composite score reduces the influence of measurement error and provides a holistic view of the underlying construct, improving the validity and precision of interpretation and aiding in understanding this construct over time. All analyses utilized the lavaan package in R (Rosseel, 2012).

Consistent with the development of the NIHTB V3, statistical goodness of fit criteria recommend that the Root Mean Square Error of Approximation (RMSEA) should have a

value of $< .06$ for a close fit and a value of $< .08$ for a reasonable fit (Browne & Cudeck, 1992; Hooper et al., 2008; Hu & Bentler, 1999; Jöreskog & Sörbom, 1993). The Standardized Root Mean Residual (SRMR) should have a value of $< .05$ for a good fit (Byrne & Russon, 1998; Hooper et al., 2008) and a value of $< .08$ for a reasonable fit (Hu & Bentler, 1999). For the Comparative Fit Index (CFI) and the Tucker Lewis Index (TLI) values, a value of > 0.95 would indicate a good fit (Hu & Bentler, 1999; Hooper et al., 2008), while values of < 0.90 - 0.95 show a reasonable fit (Kline, 2023).

4.3.2.1 Executive Function/Memory Composite

Intercorrelations in Tables 4.27 for touch-based measures for children aged 22 months and older and in Table 4.28 for measures for children aged under 22 months were consistent with a hypothesized latent variable of an Executive Function/Memory ability. Confirmatory factor analyses using Memory Task (Learning), Memory Task (Test), Visual Delayed Response, and Mullen Visual Reception scores were conducted in a single-factor model using Full Information Maximum Likelihood (FIML) to account for missing data. The final confirmatory model fit statistics showed that RMSEA was a close fit (RMSEA = .000), and the SRMR was a close fit (SRMR = .007). CFI and TLI also showed a good fit (CFI = 1.00, TLI = 1.011). The final regression weighted sum composite does not include Mullen Visual Reception for brevity, thus primarily assessing constructs of Memory and Executive Function only.

4.3.2.2 Math Composite

Intercorrelations in Table 4.29 were consistent with a hypothesized Early Math/Numeracy latent variable. Confirmatory latent factor analyses of a Math ability using all four measures (Who Has More, Subitizing, Counting, and Verbal Arithmetic scores) were conducted using a single-factor model and FIML to account for missing data. The final confirmatory model showed a close fit for RMSEA (RMSEA = .000) and the SRMR was a close fit (SRMR = .047). CFI and TLI also showed a good fit (CFI = 0.984, TLI = 0.952).

4.3.2.3 Motor Composite

Intercorrelations in Table 4.30 were consistent with a hypothesized Motor latent variable. Confirmatory latent factor analyses of a Motor ability using the Gross Motor score (Get Up and Go), Fine Motor score (Reach to Eat [Cheerio]), and Manual Dexterity score (Reach to Eat [Block]) was conducted using a single-factor model and FIML to account for missing data. The final model demonstrated a close fit via RMSEA (RMSEA = .000) and SRMR (SRMR = .007). CFI and TLI also showed a good fit (CFI = 1.00, TLI = 1.00). However, note that individuals who received a Fine Motor score could not receive a Manual Dexterity (Block) score or a Reach to Eat (Block) score, and vice versa, thus leading to a saturated model and near perfect model fit.

4.3.2.4 Mullen Language Composite

Intercorrelations in Tables 4.25 and 4.26 were consistent with a hypothesized Language latent variable. Confirmatory latent factor analyses of a Language ability were conducted using the Mullen Receptive (English and Spanish), Mullen Expressive (English and Spanish), CDI-CAT Production (English), CDI-CAT Production (Spanish), CDI-CAT Comprehension (English), CDI-CAT Comprehension (Spanish), Picture Vocabulary (English and Spanish), and Looking While Listening (English) scores. All models used FIML to account for missingness.

The final model was a single-factor model with residual covariances added between CDI-CAT Comprehension and Production in English, CDI-CAT Comprehension and Production in Spanish, and Mullen Receptive and Expressive scores to improve model fit relative to the original single-factor model. The model demonstrated a close fit for SRMR (SRMR = .066) and RMSEA (RMSEA = .005). CFI and TLI also showed a good fit (CFI = 1.00, TLI = 1.00).

4.3.2.5 Cognition Composite

A Cognition composite score was developed to enhance scoring and interpretation across all subdomains and measures within the Cognition domain. The composite score is derived from a confirmatory latent factor model using measure scores comprising the measure-level scores from the Executive Function/Memory, Math, and Mullen Language composites. In addition, the final model had residual covariances between CDI-CAT Production and Comprehension scores in English, CDI-CAT Production and Comprehension scores in Spanish, Mullen Receptive and Expressive scores, and Subitizing and Counting scores, which were added to account for commonalities in these measure scores that were not captured by the latent variables and to improve the model fit relative to the original one-factor model. The model was also fit using FIML to account for missing data. The final model demonstrated an acceptable fit via SRMR (SRMR = 0.087) but a close fit using RMSEA (RMSEA = 0.021). CFI and TLI indicated a good model fit (CFI = 0.984, TLI = 0.981).

4.3.2.6 Negative Affect Composite

The final unidimensional model for the Negative Affect composite yielded the following fit indices: CFI = 0.826, TLI = 0.816, RMSEA = 0.172, and SRMR = 0.076. The CFI and TLI values, while below the preferred threshold of 0.90, indicate a reasonable comparative fit, demonstrating that the model captures the underlying structure of Negative Affect effectively. The RMSEA of 0.172, though above the preferred threshold of 0.08, still reflects a model that is relatively close to an acceptable level of fit. Importantly, the SRMR value of 0.076 indicates a good fit, as it is below the fit cut-off of 0.08.

4.3.2.7 Self-Regulation Composite

The results of the final unidimensional model for the Self-Regulation composite are as follows: CFI = 0.930, TLI = 0.895, RMSEA = 0.159, and SRMR = 0.060. These fit indices suggest that the model demonstrates an acceptable-to-good fit, with the CFI and TLI values meeting commonly accepted thresholds, and the RMSEA and SRMR values indicating a reasonable degree of model fit.

4.3.2.8 Social Communication Composite

The final unidimensional model for the Social Communication composite yielded the following fit indices: CFI = 0.979, TLI = 0.958, RMSEA = 0.048, and SRMR = 0.024. These values indicate a strong model fit, with the CFI and TLI exceeding recommended thresholds, and the RMSEA and SRMR showing a close match to an ideal fit criterion.

4.3.3 Allowable Patterns and Composite Score Calculation

Regression weighted factor scores were calculated using Thompson's method (Thompson, 1993). Because Thompson's method requires inverting the correlation matrix among the indicators, different weights are obtained for each indicator depending on the other indicators included in the weight calculation. To account for this, when composites could be calculated using multiple theoretically justifiable combinations of indicators (see Table 4.36), weights were recalculated for each such combination. When using multiple imputation with sampling weights for weight matrix generation, the arrays required for Thompson's method (vectors of mean and SD, as well as weight matrices derived from the factor loadings and covariance matrices) were computed for each imputation, then the results were pooled across valid imputations (i.e., imputations had to converge, without any Heywood cases, and with a positive-definite latent covariance matrix) using Rubin's Rules (Rubin, 1987). At least 20 imputations were used to create pooled results.

Table 4.36

Allowable Measure Scores that Contribute to Composites for NIH Baby Toolbox Release

Composite scores	Age range (months)	Measure scores towards composite
Executive Function/Memory	22-42	Delayed Memory Task (Learning) + Delayed Memory Task (Test) + Visual Delayed Response
Math	25-42	Who Has More + Subitizing + Counting
Motor	1-42	Combination 1: Gross Motor + Manual Dexterity (Block) or

		Combination 2: Gross Motor + Fine Motor (Cheerio)
Mullen Language	1-42	Mullen Expressive + Mullen Receptive
Cognition	25-42	Measure scores from Executive Function/Memory + Math + Mullen Language composites
Negative Affect	13-42	13-36 months: ECBQ Negative Affect, PROMIS Anxiety, PROMIS Anger, PROMIS Depressive Symptoms 37-42 months: CBQ Negative Affect, PROMIS Anxiety, PROMIS Anger, PROMIS Depressive Symptoms
Self-Regulation	13-42	13-36 months: ECBQ Effortful Control, PROMIS Flexibility, PROMIS Frustration 37-42 months: CBQ Effortful Control, PROMIS Flexibility, PROMIS Frustration
Social Communication	9-23	Caregiver Checklist Younger + SOM Younger Shares Interest + SOM Younger Shares Enjoyment + SOM Younger Shares Attention + SOM Younger Play

Note. ECBQ = Early Childhood Behavior Questionnaire, CBQ = Childhood Behavior Questionnaire, SOM = Social Observation Measure.

5. References

- Aktar, E., & Pérez-Edgar, K. (2020). Infant emotion development and temperament. In J.J. Lockman, & C.S. Tamis-LeMonda (Eds.), *The Cambridge handbook of infant development*, (pp. 715–741). Cambridge University Press.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.
- Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques, second edition* (2nd ed.). CRC Press.
<https://doi.org/10.1201/9781482276725>
- Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(3), 497–529. <https://doi.org/10.1037/0033-2909.117.3.497>
- Bayley, N., & Aylward, G. P. (2019). *Bayley-4: Scales of infant and toddler development, Technical manual* (4th ed.). Pearson.
- Bergelson, E., & Swingle, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, 109(9), 3253–3258. <https://doi.org/10.1073/pnas.1113380109>
- Blackwell, C. K., Kallen, M. A., Lai, J. S., Bevans, K. B., Wakschlag, L. S., & Cella, D. (2022). Measuring PROMIS® well-being in early childhood. *Journal of Pediatric Psychology*, 47(5), 559–572. <https://doi.org/10.1093/jpepsy/jsac030>
- Blackwell, C. K., Lai, J.-S., Kallen, M., Bevans, K. B., Davis, M. M., Wakschlag, L. S., & Cella, D. (2022). Measuring PROMIS® social relationships in early childhood.

Journal of Pediatric Psychology, 47(5), 573–584.

<https://doi.org/10.1093/jpepsy/jsac031>

Blair, C., & Razza, R. P. (2007). Relating effortful control, executive function, and false belief understanding to emerging math and literacy ability in kindergarten. *Child Development*, 78(2), 647–663. <https://doi.org/10.1111/j.1467-8624.2007.01019.x>

Browne, M., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods & Research*, 21(2), 230–258.

<https://doi.org/10.1177/0049124192021002005>

Byrne, R. W., & Russon, A. E. (1998). Learning by imitation: A hierarchical approach. *Behavioral and Brain Sciences*, 21(5), 667–684.

<https://doi.org/10.1017/S0140525X98001745>

Carlson, S. M., White, R. E., & Davis-Unger, A. C. (2014). Evidence for a relation between executive function and pretense representation in preschool children. *Cognitive Development*, 29, 1–16. <https://doi.org/10.1016/j.cogdev.2013.09.001>

Cella, D., Blackwell, C. K., & Wakschlag, L. S. (2022). Bringing PROMIS to early childhood: Introduction and qualitative methods for the development of early childhood parent report instruments. *Journal of Pediatric Psychology*, 47(5), 500–509. <https://doi.org/10.1093/jpepsy/jsac027>

Clements, D. H., Sarama, J. H., & Liu, X. H. (2008). Development of a measure of early mathematics achievement using the Rasch model: The research-based early maths assessment. *Educational Psychology*, 28(4), 457–482.

<https://doi.org/10.1080/01443410701777272>

- Colombo, J. (1997). Individual differences in infant cognition: Methods, measures, and models. *Developing Brain Behaviour* (pp. 339–385). Academic Press.
<https://doi.org/10.1016/B978-012218870-1/50010-5>
- Colombo, J., & Mitchell, D. W. (2009). Infant visual habituation. *Neurobiology of Learning and Memory*, 92(2), 225–234. <https://doi.org/10.1016/j.nlm.2008.06.002>
- Colombo, J., Shaddy, D. J., Richman, W. A., Maikranz, J. M., & Blaga, O. M. (2004). The developmental course of habituation in infancy and preschool outcome. *Infancy*, 5(1), 1–38. https://doi.org/10.1207/s15327078in0501_1
- DeBell, M., & Krosnick, J. A. (2009). Computing weights for American national election study survey data. *Nes012427. Ann Arbor, MI, Palo Alto, CA: American National Election Study Technical Report Series*. <https://electionstudies.org/wp-content/uploads/2018/04/nes012427.pdf>
- Denham, S. A. (2006). Social-emotional competence as support for school readiness: What is it and how do we assess it? *Early Education and Development*, 17(1), 57–89. https://doi.org/10.1207/s15566935eed1701_4
- Denham, S. A., Wyatt, T. M., Bassett, H. H., Echeverria, D., & Knox, S. S. (2009). Assessing social-emotional development in children from a longitudinal perspective. *Journal of Epidemiology & Community Health*, 63, i37–i52.
<https://doi.org/10.1136/jech.2007.070797>
- Duncan, G. J., Dowsett, C. J., Claessens, A., Magnuson, K., Huston, A. C., Klebanov, P., Pagani, L. S., Feinstein, L., Engel, M., Brooks-Gunn, J., Sexton, H., Duckworth, K., & Japel, C. (2007). School readiness and later achievement.

- Developmental Psychology*, 43(6), 1428–1446. <https://doi.org/10.1037/0012-1649.43.6.1428>
- Eisenberg, N., Cumberland, A., Spinrad, T. L., Fabes, R. A., Shepard, S. A., Reiser, M., Murphy, B. C., Losoya, S. H., & Guthrie, I. K. (2001). The relations of regulation and emotionality to children's externalizing and internalizing problem behavior. *Child Development*, 72(4), 1112–1134. <https://doi.org/10.1111/1467-8624.00337>
- Eisenberg, N., Fabes, R. A., Guthrie, I. K., & Reiser, M. (2000). Dispositional emotionality and regulation: Their role in predicting quality of social functioning. *Journal of Personality and Social Psychology*, 78(1), 136–157. <https://doi.org/10.1037/0022-3514.78.1.136>
- Fantz, R. L. (1964). Visual experience in infants: Decreased attention to familiar patterns relative to novel ones. *Science*, 146(3644), 668–670. <https://doi.org/10.1126/science.146.3644.668>
- Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, 59(5), 1–185.
- Fernald, A., Zangl, R., Portillo, A. L., & Marchman, V. A. (2008). Looking while listening. In A. I. Sekerina, E.M. Fernandez, & H Clahsen (Ed.), *Developmental Psycholinguistics* (pp. 97–135). John Benjamins. <https://doi.org/10.1075/lald.44>
- Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, 44(3), 677–694. <https://doi.org/10.1017/S0305000916000209>

- Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The Wordbank project*. MIT Press.
- Gershon, R. C., Cook, K. F., Mungas, D., Manly, J. J., Slotkin, J., Beaumont, J. L., & Weintraub, S. (2014). Language measures of the NIH Toolbox cognition battery. *Journal of the International Neuropsychological Society*, 20(6), 642–651.
<https://doi.org/10.1017/S1355617714000411>
- Gershon, R. C., Rothrock, N., Hanrahan, R., Bass, M., & Cella, D. (2010). The use of PROMIS and assessment center to deliver patient-reported outcome measures in clinical research. *Journal of Applied Measurement*, 11(3), 304–314.
- Gershon, R. C., Wagster, M. V., Hendrie, H. C., Fox, N. A., Cook, K. F., & Nowinski, C. J. (2013). NIH Toolbox for assessment of neurological and behavioral function. *Neurology*, 80(11_supplement_3) S2–S6.
<https://doi.org/10.1212/WNL.0b013e3182872e5f>
- Gershon, R., Novack, M. A., & Kaat, A. J. (2024). The NIH Infant and Toddler Toolbox: A new standardized tool for assessing neurodevelopment in children ages 1–42 months. *Child Development*, 95(6), 2252–2254.
<https://doi.org/10.1111/cdev.14135>
- Gesell, A. (1926). A comparative method for demonstration of normal development in infancy. *Journal of the American Medical Association*, 86(17), 1277–1281.
- Ghandour, R. M., Moore, K. A., Murphy, K., Bethell, C., Jones, J. R., Harwood, R., Buerlein, J., Kogan, M., & Lu, M. (2019). School readiness among U.S. children: Development of a pilot measure. *Child Indicators Research*, 12(4), 1389–1411.
<https://doi.org/10.1007/s12187-018-9586-8>

- Hohm, E., Jennen-Steinmetz, C., Schmidt, M. H., & Laucht, M. (2007). Language development at ten months: Predictive of language outcome and school achievement ten years later? *European Child & Adolescent Psychiatry*, 16(3), 149–156. <https://doi.org/10.1007/s00787-006-0567-y>
- Hooper, D., Coughlan, J., & Mullen, M. (2008). Evaluating model fit: A synthesis of the structural equation modelling literature. *Proceedings of the 7th European Conference on Research Methodology for Business and Management Studies*, 195–200. Academic Conferences Limited.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55.
<https://doi.org/10.1080/10705519909540118>
- Jones, D. E., Greenberg, M., & Crowley, M. (2015). Early social-emotional functioning and public health: The relationship between kindergarten social competence and future wellness. *American Journal of Public Health*, 105(11), 2283–2290.
<https://doi.org/10.2105/ajph.2015.302630>
- Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific Software International.
- Kachergis, G., Marchman, V. A., Dale, P., Mankewitz, J., Frank, M. C. (2022). Online computerized adaptive tests (CAT) of children's vocabulary development in English and Mexican Spanish. *Journal of Speech Hearing and Language Research*, 65(6), 2288–2308. https://doi.org/10.1044/2022_JSLHR-21-00372

- Keen, R., Lee, M.-H., & Adolph, K. (2014). Planning an action: A developmental progression in tool use. *Ecological Psychology*, 26(1–2), 98–108.
<https://doi.org/10.1080/10407413.2014.874917>
- Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). The Guilford Press.
- LaForte, E. M., Hook, J. N., & Giella, A. K. (2024). *National Institutes of Health (NIH) Toolbox® Version 3 Technical Manual*.
- Lai, J.-S., Kallen, M. A., Blackwell, C. K., Wakschlag, L. S., & Cella, D. (2022). Psychometric considerations in developing PROMIS® measures for early childhood. *Journal of Pediatric Psychology*, 47(5), 510–522.
<https://doi.org/10.1093/jpepsy/jsac025>
- Libertus, M. E., & Brannon, E. M. (2010). Stable individual differences in number discrimination in infancy. *Developmental Science*, 13(6), 900–906.
<https://doi.org/10.1111/j.1467-7687.2009.00948.x>
- Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. *Developmental Science*, 11(3). <https://doi.org/10.1111/j.1467-7687.2008.00671.x>
- Maris, G., & van der Maas, H. (2012). Speed-accuracy response models: Scoring rules based on response time and accuracy. *Psychometrika*, 77(4), 615–633.
<https://doi.org/10.1007/s11336-012-9288-y>
- Marsala, G., & VanSant, A. F. (1998). Age-related differences in movement patterns used by toddlers to rise from a supine position to erect stance. *Physical Therapy*, 78(2), 149–159. <https://doi.org/10.1093/ptj/78.2.149>

- Mathias, S., Nayak, U. S., & Isaacs, B. (1986). Balance in elderly patients: The “get-up and go” test. *Archives of Physical Medicine and Rehabilitation*, 67(6), 387–389.
- Maydeu-Olivares, A., & Brown, A. (2010). Item response modeling of paired comparison and ranking data. *Multivariate Behavioral Research*, 45(6), 935–974.
<https://doi.org/10.1080/00273171.2010.531231>
- McCall, R. B., & Carriger, M. S. (1993). A meta-analysis of infant habituation and recognition memory performance as predictors of later IQ. *Child Development*, 64(1), 57–79. <https://doi.org/10.2307/1131437>
- McCarty, M. E., Clifton, R. K., & Collard, R. R. (1999). Problem solving in infancy: The emergence of an action plan. *Developmental Psychology*, 35(4), 1091–1102.
<https://doi.org/10.1037/0012-1649.35.4.1091>
- McCarty, M. E., Clifton, R. K., & Collard, R. R. (2001). The beginnings of tool use by infants and toddlers. *Infancy*, 2(2), 233–256.
https://doi.org/10.1207/S15327078IN0202_8
- McClelland, M. M., Cameron, C. E., Connor, C. M., Farris, C. L., Jewkes, A. M., & Morrison, F. J. (2007). Links between behavioral regulation and preschoolers’ literacy, vocabulary, and math skills. *Developmental Psychology*, 43(4), 947–959.
<https://doi.org/10.1037/0012-1649.43.4.947>
- McGraw, M. B. (1940). Neuromuscular development of the human infant as exemplified in the achievement of erect locomotion. *The Journal of Pediatrics*, 17(6), 747–771. [https://doi.org/10.1016/S0022-3476\(40\)80021-8](https://doi.org/10.1016/S0022-3476(40)80021-8)
- Moffitt, T. E., Arseneault, L., Belsky, D., Dickson, N., Hancox, R. J., Harrington, H., Houts, R., Poulton, R., Roberts, B. W., Ross, S., Sears, M. R., Thomson, W. M.,

- & Caspi, A. (2011). A gradient of childhood self-control predicts health, wealth, and public safety. *Proceedings of the National Academy of Sciences*, 108(7), 2693–2698. <https://doi.org/10.1073/pnas.1010076108>
- Mosier, C. I. (1943). On the reliability of a weighted composite. *Psychometrika*, 8(3), 161–168. <https://doi.org/10.1007/BF02288700>
- Mullen, E. M. (1995). *Mullen scales of early learning manual*. American Guidance Service.
- Navarro, M. G., Braham, E. J., & Libertus, M. E. (2018). Intergenerational associations of the approximate number system in toddlers and their parents. *British Journal of Developmental Psychology*, 36(4), 521–539. <https://doi.org/10.1111/bjdp.12234>
- Oosterhuis, H. E. M., van der Ark, L. A., & Sijtsma, K. (2016). Sample size requirements for traditional and regression-based norms. *Assessment*, 23(2), 191–202. <https://doi.org/10.1177/1073191115580638>
- Osher, D., Cantor, P., Berg, J., Steyer, L., & Rose, T. (2021). Drivers of human development: How relationships and context shape learning and development 1. In *The science of learning and development* (pp. 55–104). Routledge.
- Pelphrey, K. A., Reznick, J. S., Davis Goldman, B., Sasson, N., Morrow, J., Donahoe, A., & Hodgson, K. (2004). Development of visuospatial short-term memory in the second half of the 1st year. *Developmental Psychology*, 40(5), 836–851. <https://doi.org/10.1037/0012-1649.40.5.836>

- Podsiadlo, D., & Richardson, S. (1991). The timed “up & go”: A test of basic functional mobility for frail elderly persons. *Journal of the American Geriatrics Society*, 39(2), 142–148. <https://doi.org/10.1111/j.1532-5415.1991.tb01616.x>
- Putnam, S. P., Gartstein, M. A., & Rothbart, M. K. (2006). Measurement of fine-grained aspects of toddler temperament: The early childhood behavior questionnaire. *Infant Behavior and Development*, 29(3), 386–401. <https://doi.org/10.1016/j.infbeh.2006.01.004>
- Putnam, S. P., Helbig, A. L., Gartstein, M. A., Rothbart, M. K., & Leerkes, E. (2014). Development and assessment of short and very short forms of the infant behavior questionnaire–revised. *Journal of Personality Assessment*, 96(4), 445–458. <https://doi.org/10.1080/00223891.2013.841171>
- Putnam, S. P., & Rothbart, M. K. (2006). Development of short and very short forms of the children’s behavior questionnaire. *Journal of Personality Assessment*, 87(1), 102–112. https://doi.org/10.1207/s15327752jpa8701_09
- Putnam, S. P., Rothbart, M. K., & Gartstein, M. A. (2008). Homotypic and heterotypic continuity of fine-grained temperament during infancy, toddlerhood, and early childhood. *Infant and Child Development*, 17(4), 387–405. <https://doi.org/10.1002/icd.582>
- Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, 47(5), 667–696. <https://doi.org/10.1080/00273171.2012.715555>
- Revelle, W. (2015). Package ‘psych.’ *The Comprehensive R Archive Network*, 337(338), 161–165.

- Revelle, W., & Condon, D. M. (2019). Reliability from α to ω : A tutorial. *Psychological Assessment*, 31(12), 1395–1411. <https://doi.org/10.1037/pas0000754>
- Rine, R. M., Schubert, M. C., Whitney, S. L., Roberts, D., Redfern, M. S., Musolino, M. C., Roche, J. L., Steed, D. P., Corbin, B., Lin, C.-C., Marchetti, G. F., Beaumont, J., Carey, J. P., Shepard, N. P., Jacobson, G. P., Wrisley, D. M., Hoffman, H. J., Furman, G., & Slotkin, J. (2013). Vestibular function assessment using the NIH Toolbox. *Neurology*, 80(11_supplement_3), S25–S31. <https://doi.org/10.1212/WNL.0b013e3182872c6a>
- Rose, S. A., Feldman, J. F., Wallace, I. F., & Cohen, P. (1991). Language: A partial link between infant attention and later intelligence. *Developmental Psychology*, 27(5), 798–805. <https://doi.org/10.1037/0012-1649.27.5.798>
- Rose-Krasnor, L. (1997). The nature of social competence: A theoretical review. *Social Development*, 6(1), 111–135. <https://doi.org/10.1111/j.1467-9507.1997.tb00097.x>
- Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48(2), 1–36. <https://doi.org/10.18637/jss.v048.i02>
- Rothbart, M. K., Ahadi, S. A., & Hershey, K. L. (1994). Temperament and social behavior in childhood. *Merrill-Palmer Quarterly* (40)1, 21–39. <http://www.jstor.org/stable/23087906>
- Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.
- Sacrey, L.-A. R., Karl, J. M., & Whishaw, I. Q. (2012). Development of rotational movements, hand shaping, and accuracy in advance and withdrawal for the

- reach-to-eat movement in human infants aged 6–12 months. *Infant Behavior and Development*, 35(3), 543–560. <https://doi.org/10.1016/j.infbeh.2012.05.006>
- Salsman, J. M., Lai, J.-S., Hendrie, H. C., Butt, Z., Zill, N., Pilkonis, P. A., Peterson, C., Stoney, C. M., Brouwers, P., & Cella, D. (2014). Assessing psychological well-being: Self-report instruments for the NIH Toolbox. *Quality of Life Research*, 23(1), 205–215. <https://doi.org/10.1007/s11136-013-0452-3>
- Sherlock, P., Blackwell, C. K., Kallen, M. A., Lai, J.-S., Cella, D., Krogh-Jespersen, S., Luby, J. L., Buss, K. A., Burns, J., & Wakschlag, L. S. (2022). Measuring PROMIS® emotional distress in early childhood. *Journal of Pediatric Psychology*, 47(5), 547–558. <https://doi.org/10.1093/jpepsy/jsac029>
- Squires, J., Bricker, D., & Potter, L. (1997). Revision of a parent-completed developmental screening tool: ages and stages questionnaires. *Journal of Pediatric Psychology*, 22(3), 313–328. <https://doi.org/10.1093/jpepsy/22.3.313>
- Squires, J., & Bricker, D. (2009). *Ages and stages questionnaires* (3rd ed.). Paul H. Brookes. <http://www.bestkc.com/wp-content/uploads/2017/05/3-year-ASQ-ENGLISH.pdf>
- Starr, A., Libertus, M. E., & Brannon, E. M. (2013). Number sense in infancy predicts mathematical abilities in childhood. *Proceedings of the National Academy of Sciences*, 110(45), 18116–18120. <https://doi.org/10.1073/pnas.1302751110>
- Stasinopoulos, M. D., Rigby, R. A., & Bastiani, F. D. (2018). GAMLSS: A distributional regression approach. *Statistical Modelling*, 18(3–4), 248–273. <https://doi.org/10.1177/1471082X18759144>

- Stauffer, J. M., & Mendoza, J. L. (2001). The proper sequence for correcting correlation coefficients for range restriction and unreliability. *Psychometrika*, 66(1), 63–68.
<https://doi.org/10.1007/BF02295732>
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217), 1632–1634.
<https://doi.org/10.1126/science.7777863>
- Thompson, B. (1993). Calculation of standardized, noncentered factor scores: An alternative to conventional factor scores. *Perceptual and Motor Skills*, 77(3_suppl), 1128–1130. <https://doi.org/10.2466/pms.1993.77.3f.1128>
- Thorndike, R. L. (1949). *Personnel selection; test and measurement techniques*. Wiley
<https://psycnet.apa.org/record/1949-05074-000>
- U.S. Census Bureau. (2019). *American community survey 5-year estimates*. Retrieved from <https://www.census.gov/data/developers/data-sets/acs-5year.2019.html>
- U.S. Census Bureau. (2022). *American community survey 5-year estimates*. Retrieved from <https://www.census.gov/data/developers/data-sets/acs-5year.2022.html>
- van der Maas, H. L., & Wagenmakers, E.-J. (2005). A psychometric analysis of chess expertise. *The American Journal of Psychology*, 118(1), 29–60.
- Weintraub, S., Dikmen, S. S., Heaton, R. K., Tulsky, D. S., Zelazo, P. D., Bauer, P. J., Carlozzi, N. E., Slotkin, J., Blitz, D., Wallner-Allen, K., Fox, N. A., Beaumont, J. L., Mungas, D., Nowinski, C. J., Richler, J., Deocampo, J. A., Anderson, J. E., Manly, J. J., Borosh, B., Havlik, R., Conway, K., et al. (2013). Cognition

assessment using the NIH Toolbox. *Neurology*, 80(11 Suppl 3), S54–S64.

<https://doi.org/10.1212/WNL.0b013e3182872ded>

Wetherby, A. M., Brosnan-Maddox, S., Peace, V., & Newton, L. (2008). Validation of the Infant-Toddler Checklist as a broadband screener for autism spectrum disorders from 9 to 24 months of age. *Autism*, 12(5), 487–511.

<https://doi.org/10.1177/1362361308094501>