Supplemental Document A

Bayley Scales of Infant and Toddler Development™, Fourth Edition (Bayley™-4) is a standard series of measurements originally created to assess the development of infants and toddlers, ages 16 days–42 months. This measure consists of a series of developmental play tasks and takes 30-70 minutes to administer. Subtest level scaled scores, domain level composite scores, percentile ranks, confidence intervals, developmental age equivalents, and growth scale values are provided. Growth scale value (GSVs), provide a score metric that does not involve age-based comparisons. GSVs have a mean of 500 and SD of 100. The Bayley-4 assesses development across five scales: Cognitive, Language, Motor, Social-Emotional, and Adaptive Behavior. Assessment of the Cognitive, Language, and Motor domains is accomplished by direct assessment. The Social-Emotional and Adaptive Behavior scales are administered via caregiver questionnaire. Reliability estimates for the normative sample for the scales are: Cognitive .95, Language .96, Motor, .95, Social-Emotional .91, and .98 for the Adaptive Behavior Scale.

Adaptive Behavior Assessment System, Third Edition (ABAS-3) is a standardized measure of adaptive behavior developed for individuals from birth through 89 years of age. The ABAS-3 covers three broad adaptive domains: Conceptual, Social, and Practical. Within these domains, it assesses 11 adaptive skill areas (each form assesses 9 or 10 skill areas based on age range). Items focus on practical, everyday activities required to function, meet environmental demands, care for oneself, and interact with others effectively and independently. On a 4-point response scale, parents to caregivers indicate whether the individual can perform each activity, and if so, how frequently they perform it when needed. Higher scores indicate greater competence. Internal consistency reliability estimates range from .80 to .99, and inter-rater reliability estimates range from .70 to .91. It has good to excellent test-retest reliability.
**Infant Behavior Questionnaire-Revised (IBQ-R)**\(^{39}\) is a 191-item measure, organized into 14 subscales (activity level, distress to limitations, fear, duration of orienting, smile/laughter, high pleasure, low pleasure, soothability, falling reactivity, cuddliness, perceptual sensitivity, sadness, approach, and vocal reactivity), that is designed to assess infant temperament between the ages of 3 and 12 months. The frequency of infant behaviors are rated on a scale from 1 (never) to 7 (always), with higher scores indicating more of each construct. Internal reliability has previously been demonstrated to be good for each of the subscales.\(^{39}\) Conceptual and item analyses provided support for 14 of the 16 proposed scales, demonstrating satisfactory internal consistency. Inter-rater reliability was also evaluated, with evidence of moderate agreement between primary and secondary caregivers. Mono-method discriminant validity was demonstrated through an examination of correlations among the IBQ-R scale scores. Results of the factor analytic procedure were consistent with three broad dimensions of Surgency/Extraversion, Negative Affectivity, and Orienting/Regulation.

**Brief Infant Sleep Questionnaire (BISQ)**\(^{40}\) is a 13-item parent report questionnaire (9 sleep-related questions, 4 demographic questions) that screens for infant and toddler sleep problems for children ages 0-3 years. Questions are multiple-choice or open-ended (e.g. How much time does your child spend in sleep during the day?). Clinical cutoffs for referral are suggested for the following behaviors: wakes >3 times per night, spends > 1 hour of wakefulness in the night, spends < 9 hours of sleep (day and night). Test-retest coefficients range from .82 to .95. Significant relationships between the BSIQ, actigraphic, and sleep log measures have also been reported, and it has been found to correlate with rated severity of sleep problems.

**Brief Infant-Toddler Social and Emotional Assessment (BITSEA)**\(^{41}\) is a 42-item questionnaire designed to assess for social-emotional and behavior problems in young children aged 12-36
months old according to caregiver report. Items are rated on a 3-point scale and compose a problems scale (31 items) and a competence scale (11 items); higher scores indicate more problems and greater competence, respectively. Good test-retest reliability has been reported by the developers of the measure; the BITSEA also demonstrated high correlations with the ITSEA (longer version of the BITSEA) and other relevant measures. In a higher-risk, community sample, the BITSEA total problem scale was also shown to have high internal consistency and concurrent validity.

_Ages & Stages Questionnaire-3 (ASQ-3)²¹_ is a method to screen children for developmental delays between 1 month and 5½ years, without any gaps between the questionnaire age intervals. Parents/caregiver answer questions on a Likert scale (i.e. yes, sometimes, not yet) about their infant/child’s fine motor, gross motor, communication, problem-solving and personal-social development. There are established cut-offs for each domain, which signify whether a child should be referred for comprehensive assessment or monitored. Research with a sample of 15,138 diverse children show that the ASQ-3 is reliable and valid. ASQ-3 identifies children for further assessment with excellent sensitivity (.86) and specificity (.85).

_Comunication and Symbolic Behavior Scales Developmental Profile™ (CSBS DP™) Infant-Toddler Checklist_⁴³ is used to identify early delays in social communication, expressive speech/language, and symbolic functioning. Parents/caregivers respond to 24 multiple choice questions about their child’s communication. It is a norm-referenced screening and evaluation tool that helps determine the communicative competence of infants and toddlers, measured by seven language predictors: emotion and eye gaze, communication, gestures, sounds, words, understanding, and object use. From these seven clusters, three composites are generated (Social, Speech, Symbolic). Standard scores and composite percentiles are generated. Higher scores reveal better performance. CSBS DP is used with infants and toddlers whose functional
communication age is between 6 months and 24 months. CSBS DP was successfully tested for validity and reliability with large samples of children (2,188 for the checklist, 790 for the Caregiver Questionnaire, and 337 for the Behavior Sample).

**M-CHAT-R/F** is a two-step screening tool used to detect children likely to have autism. It is intended for use at regular well-child checkups for children 16 to 30 months old. With the M-CHAT-R/F, health care providers can classify a child’s risk of having autism as low, medium, or high, on the basis of parents’ answers to 20 questions. A score in the high-risk range warrants a referral for further evaluation for possible autism. For a child determined to be at medium risk, M-CHAT R/F includes a follow-up questionnaire used soon after the original evaluation to obtain additional information needed to more definitively classify the child as either high risk or low risk. The reliability and validity of the M-CHAT-R/F were demonstrated, and optimal scoring was determined by using receiver operating characteristic curves. Children whose total score was ≥3 initially and ≥2 after follow-up had a 47.5% risk of being diagnosed with ASD (confidence interval [95% CI]: 0.41–0.54) and a 94.6% risk of any developmental delay or concern (95% CI: 0.92–0.98). Total score was more effective than alternative scores. An algorithm based on three risk levels is recommended to maximize clinical utility and to reduce age of diagnosis and onset of early intervention.

**Edinburgh Postnatal Depression Scale (EPDS)** is a 10-item questionnaire that was developed to identify women who have prenatal and postpartum depression symptoms. Items on the scale correspond to various clinical depression symptoms, such as guilt feelings, sleep disturbance, low energy, anhedonia, and suicidal ideation. The overall score is determined by adding together the scores for each of the 10 items, with some items reverse scored. Higher scores indicate more depressive symptoms. The EPDS is described as having “satisfactory validity,” in that it had been validated across these two samples. Seven items were constructed by the
Researchers while six others were adapted from the Irritability Depression and Anxiety Scale (IDA) and the Hospital Anxiety and Depression Scale (HAD). In considering reliability, Cox and colleagues found that the scale had a split-half reliability of 0.88 and the standardized a coefficient was 0.87. The EPDS was correlated with the Postpartum Depression Screening Scale (PDSS) \( r = .79 \), providing evidence for convergent validity. Within our sample, the EPDS demonstrated good internal reliability at T1 (\( \alpha = .76 \)) and at T2 (\( \alpha = .84 \)).

**Brief Symptom Inventory (BSI)** is a 53-item self-report inventory in which participants rate the extent to which they have been bothered (0 = "not at all" to 4 = "extremely") in the past week by various symptoms. The BSI has nine subscales designed to assess individual symptom groups: somatization, obsessive-compulsive, interpersonal sensitivity, depression, anxiety, hostility, phobic anxiety, paranoia, and psychosocial. Higher scores indicate higher levels of each construct. The BSI also includes three scales that capture global psychological distress. Multiple subscale scores are provided including a Global Severity Index (GSI), Positive Symptoms Total, and Positive Symptoms Distress Index. Reliability for the GSI Total is reported as .95.

**Postpartum Bonding Questionnaire (PBQ)** has 25 statements, each rated along a 6-point scale. Positive responses, such as "I enjoy playing with my baby," are scored from 0 ("always") to 5 ("never"). Negative responses, such as "I am afraid of my baby," are scored from 5 ("always") to 0 ("never"). Response patterns yield scores on four factors: 1) general factor (Scale 1), rejection and pathological anger (Scale 2), anxiety about the infant (Scale 3), and incipient abuse (Scale 4), with higher scores indicating more pathology. Validation of the PBQ revealed sensitivity values of .82 for Scale 1 (general factor), .88 for Scale 2 (rejection), and .67 for Scale 2 (pathological anger).
Parenting Stress Index-4-Short Form (PSI-4)\textsuperscript{47} is an abbreviated version of the full-length, parent self-report, PSI-4. Thirty-six items are divided into three domains: Parental Distress (PD), Parent-Child Dysfunctional Interaction (P-CDI), and Difficult Child (DC), which combine to form a Total Parenting Stress scale. Higher scores on all totals indicate more parenting stress. This scale is designed for use with parents of children 3 months to 12 years old. Coefficient alphas for each PSI-4-SF scale are all above .90.

Maternal Self-Report Inventory-Short Form (MSI-SF)\textsuperscript{48} is a 26-item questionnaire measuring maternal self-esteem in the areas of: caretaking ability, general ability and preparedness for mothering role, acceptance of baby, expected relationship with baby, and feelings concerning pregnancy, labor, and delivery. Higher scores indicate greater feelings of maternal efficacy/esteem. The scale has adequate internal consistency, with alpha coefficients in the .80s, and has been associated with other measures of general self-efficacy.

Inventory of Problems-29 (IOP-29)\textsuperscript{67} is a brief, self-administered, paper-and-pencil, measure of feigned cognitive and mental disorders. Research supports the utility of the IOP-29 for discriminating valid from feigned cognitive and psychiatric complaints. Classifications within clinical samples with base rates of .5 produced sensitivity, specificity, positive predictive power, and negative predictive power statistics of .80. The IOP-29 will serve as one of our external criterion measures in the development of our embedded validity indices.

Personality Assessment Inventory (PAI)\textsuperscript{68} is an objective, 344 item inventory of adult personality that assesses psychopathological syndromes and provides information relevant for clinical diagnosis, treatment planning, and screening for psychopathology. It has 22 nonoverlapping scales including four validity scales: 1) Inconsistency (INC), 2) Infrequency (INF), 3) Negative Impression (NIM), and 4) Positive Impression (PIM). The INF scale is useful in identifying
individuals who complete the PAI in atypical ways sue to random responding, indifference, carelessness, confusion, or reading difficulties. The NIM scale contains items which present an exaggerated unfavorable impression and unlikely symptoms. The PIM scale items involve the presentation of a very favorable impression. The INF, NIM, and PIM scales were used in the development of the Modified PAI Validity Scale-42 which serves as one of our external criterion in the development of our embedded validity indices. The PAI has high internal consistency, with a median alpha for the full scale of .81 for the normative sample.