Beyond the Bayley: Neurocognitive Assessments of Development
during Infancy and Toddlerhood

Natalie H. Brito¹, William P. Fifer², Dima Amso³, Rachel Barr⁴, Martha Ann Bell⁵, Susan
Calkins⁶, Albert Flynn⁷, Hawley E. Montgomery-Downs⁸, Lisa M. Oakes⁹, John E. Richards¹⁰,
Larissa M. Samuelson¹¹, & John Colombo¹²

¹ Department of Applied Psychology, New York University, New York, NY, 10003
² Division of Developmental Neuroscience, New York State Psychiatric Institute, New York, NY
10032
³ Department of Cognitive, Linguistic, and Psychological Sciences, Brown University,
Providence, RI, 02912
⁴ Department of Psychology, Georgetown University, Washington, DC, 20057
⁵ Department of Psychology, Virginia Tech, Blacksburg, VA, 24061
⁶ Department of Human Development and Family Studies, University of North Carolina at
Greensboro, Greensboro, NC, 27412
⁷ School of Food and Nutritional Sciences, University College Cork, Cork, Ireland
⁸ Department of Psychology, West Virginia University, Morgantown, WV, 26506
⁹ Department of Psychology, University of California, Davis, Davis, CA, 95616
¹⁰ Department of Psychology, University of South Carolina, Columbia, SC, 29208
¹¹ Department of Psychology, University of East Anglia, Norfolk, United Kingdom
¹² Department of Psychology, University of Kansas, Lawrence, KS, 66045

Key words: cognition; infants; toddlers; measurement

Word Count: 11,120

Corresponding Author:
Natalie Hiromi Brito
New York University
Kimball Hall 407W
246 Greene Street
New York, NY 10003
natalie.brito@nyu.edu
ABSTRACT

The use of global, standardized instruments is conventional among clinicians and researchers interested in assessing neurocognitive development. Exclusively relying on these tests for evaluating effects may underestimate or miss specific effects on early cognition. The goal of this review is to identify alternative measures for possible inclusion in future clinical trials and interventions evaluating early neurocognitive development. The domains included for consideration are attention, memory, executive function, language and socio-emotional development. Although domain-based tests are limited, as psychometric properties have not yet been well-established, this review includes tasks and paradigms that have been reliably used across various developmental psychology laboratories.
Introduction

Development occurs at a rapid pace during the first three years of life, with significant changes taking place in cognitive, language, and social skills during toddlerhood. The quality of both the learning environment and pro-social relationships contribute to a child’s developmental trajectory and are the foundation for subsequent learning and development. Cognitive skills measured in early childhood show increasing levels of stability (Carlson, 2005; Posner & Rothbart, 2000) and associations with later academic outcomes (Alloway & Alloway, 2010; Fagan, Holland, & Wheeler, 2007). Therefore, accurate assessments of early neurodevelopment, particularly during toddlerhood, provide important measures of concurrent and future cognitive functioning. Compared to school-aged children, very young children are hard to assess reliably and validly due to limits in their motor, language, and socio-emotional skills. Young children also show great variability within and across individuals, potentially reflecting the emerging differentiation of functional systems (Karmiloff-Smith, 2012).

The state of neurocognitive assessment of children up to 3 years of age has been neglected and inconsistently-addressed in the current scientific literature. There is a need to build consensus around reliable tasks and best practices for such assessments in order to evaluate the efficacy of various early interventions. Agreement from the developmental science community would permit more domain-specific measures to be utilized during clinical research trials. The purpose of this review is to identify candidate measures of a possible standard “toolkit” for inclusion in future clinical trials and interventions evaluating the effects of early neurocognitive development. Not all domains are covered and as the field of developmental psychology contains numerous potential candidates for inclusion, we focus on tasks and paradigms that have been reliably used within our own respective laboratories. The domains included for consideration in
this age range (0-3) are attention, memory, executive function, language and socio-emotional development. Sections include discussions across different levels of measurement (e.g., behavior, electrophysiology), with a concluding section devoted to additional variables of interest that may help to contextualize early neurocognitive findings. First, a brief overview of the most commonly used global assessment for early neurodevelopment, as well as example of a collection of domain-specific tasks for older children is presented.

**Global Assessments of Early Neurodevelopment**

Most clinical studies or interventions have generally relied on global, standardized tests for evaluations of early learning and development. These tests are derived from normative developmental milestones, and generally assume the presence of a unique underlying factor which drives all individual differences in mental or behavioral performance (Uzgiris & Hunt, 1975). The implication here is that cognitive or mental status may be adequately quantified in terms of a single, overall composite score. These tests are advantageous in many ways. They are well-standardized, have established psychometric properties, are easily interpreted, and are generally well-known and widely-accepted by clinicians, pediatricians, and health practitioners.

Although there are many global assessments of cognitive functioning during toddlerhood, the most widely used test of general neurodevelopment is the Bayley Scales of Infant and Toddler Development. This assessment, designed for children ages 1 month to 42 months, has been administered in both clinical and research settings. The scales are intended to assess three major areas of development: cognitive, language and motor (Bayley, 1969; 2006). Although currently the most used option for assessments of general cognitive development, the Bayley scales are not without their limitations and criticisms. The Bayley is a fairly blunt instrument and may not pick up subtle deficits; studies have reported that the Bayley-III underestimates
developmental delay (Anderson, De Luca, Hutchinson, Roberts, & Doyle, 2010; Moore, Johnson, Haider, Hennessy, & Marlow, 2012). Although the Bayley-III is designed to assess developmental delay, it is often used by researchers to predict individual differences in cognitive functioning. In addition, it is difficult to assess specific skills that may be particularly relevant to a certain treatment or intervention. For example, if the goals of a nutrition intervention were to improve infant sleep and subsequent memory performance, no specific scores for memory could be ascertained from this assessment and any effects specific to memory may be diminished by the infant’s poorer performance in other related cognitive skills.

Specific Assessments of Early Neurodevelopment

Alternative conceptualizations of human cognition are based on process-based models derived from information-processing theory (Neisser, 1969). These assume that various cognitive processes are complexly determined and largely independent of one another. This assumption is supported by advances in neuroscience that show, for example, that cognitive functions like attention (Petersen & Posner, 2012; Webster & Ungerleider, 1998) and memory (Squire & Zola, 1996) are served by multiple functional networks. The disadvantages of process measures are often the opposite of the strengths of standardized tests: they are not standardized, often have unknown psychometric properties, are not easily interpreted, and they are not well known outside of basic developmental science communities in which they have been used. However, one particular advantage these measures may have over more global measures of cognitive status is that they may be sensitive to more specific or to subtler delays or deficits in development.

Currently, there is no collection of specific tasks or assessments to evaluate early neurodevelopmental skills during toddlerhood. An assessment package designed to test specific cognitive skills in older children and adults has been developed by the NIH, and this assessment
tool may serve as an example for future development of tasks specific to toddlerhood. The NIH Toolbox for the Assessment of Neurological and Behavioral Function was developed as part of the NIH Blueprint for Neuroscience Research in order to design a set of state-of-the-art measurement tools to enhance data collection in large cohort studies; the NIH Toolbox is not intended for use as a diagnostic tool. It assesses four major domain areas: cognitive, emotional, motor, and sensory health. Each domain is composed of multiple subdomains, which are functional constructs that are measurable representations of such (Gershon, et al., 2013; Salsman et al., 2013; Weintraub et al., 2013; Zelazo et al., 2013).

The NIH Toolbox Early Childhood Cognition Battery was designed to provide a brief, efficient computerized test of key neuropsychological function for young children ages 3 to 6 years. The tests include: Dimensional Change Card Sort (cognitive flexibility), Flanker Inhibitory Control and Attention, Pattern Comparison Processing Speed, Picture Sequence Memory (episodic memory), and the Picture Vocabulary (receptive vocabulary and language). In addition to individual task scores, the battery will yield an Early Childhood Composite Score as a general measure of early cognitive function (Weintraub et al., 2013; Zelazo et al., 2013). The NIH Toolbox is not, however, without its limitations. As part of a validation study, the Cognition Battery was administered to a sample of 1,020 typically developing children (ages 3 years to 20 years) tested at 9 sites across the U.S. (Pediatric Imaging Neurocognition and Genetics: PING, Akshoomoff, et al., 2014). The researchers observed some ceiling effects in older children and some floor effects on the executive function tasks in the younger participants. Specifically, the NIH Toolbox version of the DCCS was reported to be significantly limited in its use for measuring cognitive flexibility in children under the age of 7 and a small percentage of children were unable to meet the practice trial criteria for the Flanker Inhibitory Control and Attention
Test (Akshoomoff, et al., 2014). As all NIH Toolbox assessments are administered using a computer or touchscreen monitor, the child’s proficiency in media use must be taken into consideration when evaluating reaction time or attention performance.

Need for Alternative Assessments of Early Neurocognitive Development

Reliance on global measures of neurocognitive development can significantly impact the results and interpretation of studies that may play a role in decisions surrounding intervention efficacy or policy. For example, within clinical studies of nutrition, examining the impact of prenatal teratogens, infants exposed prenatally to alcohol show deficits in visual attention, but are not impaired in memory; conversely, infants exposed prenatally to polychlorinated biphenyls (PCBs) show deficits in memory but not in attention (Jacobson, Fein, Jacobson, Schwartz, & Dowler, 1985; Jacobson, Jacobson, Sokol, Martier, & Ager; 1993). Thus, if different compounds affect different cognitive systems, outcome measures will need to be selected carefully, and global tests may obscure specific delays or deficits. Indeed, a similar profile has emerged for the effect of various long-chain polyunsaturated fatty acids (LCPUFA) on early brain and behavioral development. LCPUFA are obtained from the diet, accumulates in brain tissue early in development, and have been widely thought to have potential beneficial effects on cognitive development. However, based on results from the Bayley Scales of Infant Development administered at 18 months of age, four recent meta-analyses on LCPUFA have concluded that there is no cognitive benefit of LCPUFA (particularly docosahexaenoic acid, or DHA) supplementation (Beyerlein et al, 2010; Simmer, Patole, & Rao, 2011; Qawasmi, Landeros-Weisenberger, Leckman, & Bloch, 2012). This overarching negative conclusion has persisted, despite a number of positive findings suggesting specific benefits of LCPUFA on visual attention in infants and children (Carlson & Werkman, 1996; Colombo et al., 2004a; Colombo, et al.,
2011; Kannass, Colombo, & Carlson, 2009; McNamara et al., 2010; Vaisman et al., 2008; Werkman & Carlson, 1996; Westerberg et al., 2010) and early manifestations of executive function (Drover, Hoffman, Castaneda, Morale, & Birch, 2009; Henriksen et al., 2008; Judge, Harel, & Lammi-Keefe, 2007; Willatts, Forsyth, DiModugno, Varma, & Colvin, 1998). The following sections outline domain-specific tasks and assessments across different levels of measurement within the areas of attention, memory, executive function, language and socio-emotional development.

**Domain-Specific Outcomes**

**Attention**

Attention plays a central role in learning and the acquisition of information. Individual differences in early attention are posited to reflect the speed and efficiency of information processing and these differences have been found to be related to later cognitive abilities (Colombo, 1993; Cuevas & Bell, 2014; Rose & Feldman, 1997). Attention is not a single unified construct, however, and assessments tap different aspects of attention. For example, some measures may reflect selection and aspects of attention related to information intake and processing. Other measures may reflect attentional control, or the ability to both use cues to guide attention and to inhibit irrelevant information. Nevertheless, given the ease of administering such tasks, and the large literature providing deep understanding into attentional processes in adults, attention outcome measures are extremely promising for the future of assessment of neurocognitive development.

**Looking.** Looking time and visual attention have been studied in infants since the 1950s. Visual behavior involves selection of some inputs and inhibition of others and may, therefore, provide broad insight into cognitive development (Colombo, 2001; Oakes, 2017). It also
involves integration of multiple pieces of information and memory systems (Colombo, 2001; Colombo & Cheatham, 2006; Posner & Rothbart, 2007). Visual behavior reflects speed of processing (e.g., Messinger, Mattson, Todd, Gangi, Myers, & Bahrick, 2017; Rose, Feldman, & Jankowski, 2002) and development of multiple brain systems (Colombo, 2001; Johnson & Vecera, 1996). Visual attention also can be measured in toddlers (Colombo, 2001), making it especially useful for assessing development over the first three years. Moreover, looking behavior is sensitive to developmental differences. In a longitudinal sample of full-term and preterm infants, results demonstrated more efficient patterns of attention for the full-term infants (i.e., shorter look durations, faster shift rates) than infants in the preterm group (Rose, Feldman, & Jankowski, 2001). In a separate line of work, infants who later develop autism show different visual scanning of faces in early infancy (Jones & Klin, 2013). Toddlers diagnosed with Autism Spectrum Disorder, Fragile X, or Williams Syndrome differ from typically developing children in their scanning of social and non-social stimuli, visual search, and disengagement (Kaldy, Kraper, Carter, & Blaser, 2011; Scerif et al., 2004; 2005). These studies demonstrate how eye-tracking measures (i.e., fine-grained measures of where children/toddlers look, how long they look, changes in where they look, how fast children look, and what makes children look at one thing versus another), can provide significant insight into developmental and individual differences in visual attention. Importantly, eye-tracking systems are increasingly portable and can be used in clinics, for “off-site” recording (Ballieux, et al., 2015).

**Heart-Rate Defined Phases of Attention.** A difficulty inherent in the use of visual behavior as a predictive measure is that not all looking reflects active visual cognition in the human infant. Respiratory and cardiac measures during looking have revealed that the degree of infants' active processing varies systematically within a look (Richards, 1985; 1987; Richards &
Ca
ey, 1990; 1991). In fact, looks may be parsed into three different "phases" of visual attention: Orienting, Sustained Attention (SA), and Attention Termination. Of these phases, SA (a period of looking characterized by cardiac deceleration) is the one that reflects active stimulus processing. Specifically, SA, but not the other phases, correlates with autonomic and behavioral measures that predict concurrent or lagged cognitive status from infancy, such as vagal tone (Linnemeyer & Porges, 1986) and successful recognition memory performance (Richards & Casey, 1990). The proportion of SA has been shown to decline over the first year (Colombo, Shaddy, Richman, Maikranz, & Blaga, 2004b), but maintenance of high levels of SA over that first year has also been reported to be associated with better language and cognitive outcomes in later childhood (Colombo, et al., 2009). Heart rate changes occurring during stimulus presentation in infant participants represent an effective measure of the alerting and attentional systems of the human brain. Stimulus processing occurs primarily during sustained attention, so that the precision of measures of infant cognitive processing could be supplemented by knowing when in a stimulus presentation sustained attention could be occurring.

**Cue Competition Paradigms.** The Infants Orienting with Attention (IOWA) task is an attentional cueing task (Ross-Sheehy, Schneegans, & Spencer, 2015) that takes advantage of the fact that between 5 and 10 months of age infants increasingly are able to use a cue (in this case a briefly presented black dot in one of two locations) to control attention. In this task, a cue is presented and then after a brief delay a target (in this case a photograph of a real object) is presented either in the same location as the target or on the opposite side of the fixation. Infants, like adults, make faster eye-movements toward the target when it appears in the cued location rather than the non-cued location. This task is sensitive and engaging; although this task has not been tested in toddlers, it has good potential.
A second alternative task for toddlers is visual search. Versions of this task have been used with toddlers in studies of typical and atypical development (Gerhardstein, Adler, & Rovee-Collier, 2000; Kaldy et al., 2011; Scerif et al., 2004, 2005). In general, visual search measures selective attention, or the ability to balance the pull of competing stimuli on visual attention, as well as the inhibition of responses to distractors. Research has linked performance on this task to learning and perception in infants (e.g., Amso & Johnson, 2006). In toddlers, visual search tasks have successfully uncovered differences between atypically and typically developing children (Kaldy et al., 2011; Scerif et al., 2004, 2005).

**Memory**

Learning and memory are inextricably linked; tests of learning are, in fact, tests of memory. Memory is the product of a series of learning processes that include encoding, storage, and retrieval. Although infants are often required to recall information under the same conditions in which they encountered it, termed memory recall, toddlers are also faced with the challenge of learning about their world from a variety of sources and must then apply what they learned to diverse problems. This ability to retrieve memories despite changes in perceptual cues, allowing learning to be generalized to novel situations, has been referred to as memory flexibility (Eichenbaum, 1997; Hayne, 2006; Karmiloff-Smith, 1994). Memory recall and flexibility profiles emerge gradually during development; early in development, successful memory performance is contingent on an exact match between the cues at the time of encoding and the cues available at retrieval. A mismatch at learning and test can decrease memory performance, but with age toddlers can increasingly tolerate differences between conditions at encoding and retrieval (for a review see Barr & Brito, 2014).
Although there are many established methods for measuring memory (e.g., novelty preference memory tests), the focus here is on operant conditioning, visual recognition memory (VRM), and deferred imitation, as these have the strongest empirical base and are the best predictors of cognitive outcomes (see Rovee-Collier & Barr, 2001; 2010). Among the principles employed to illustrate best practices for intervention tracking are 1) parameterization of tasks essential for reliable and valid assessment, 2) measurement of both memory recall and memory flexibility; 3) assessment of the predictive validity of the measures, and; 4) examination of memory as a possible precursor to other more complex cognitive outcomes.

**Operant Conditioning.** Over the past 40 years, operant conditioning has been studied utilizing the mobile conjugate reinforcement and train tasks, resulting in an extensive empirical base (Rovee-Collier & Barr, 2001). In this task, 2 to 6-month-old infants learn to kick to move an overhead mobile. After obtaining a baseline measure of foot kicking, the infant’s foot is tied to the mobile and the infant learns the contingency between foot kicking and mobile movement; long-term retention can then be assessed after a delay. During the long-term test, the infant is placed in the crib without the foot tied to the mobile and the rate of kicking is noted. Retention is assessed with a baseline ratio (test kick rate divided by the baseline kick-rate); ratios above 1.5 indicate retention. Forgetting is assessed with a retention ratio (long-term retention kick rate divided by the immediate retention rate). In order to allow for rapid changes in speed of learning, motor development, and motivation levels, baseline and training periods become shorter as infants and toddlers get older. To assess the same constructs in older children a “train task”, has been used successfully with 6- to 18-month-olds; here, infants use their hands to press a lever to make a train move around a toy train track. Memory performance on operant tasks are predictive of later cognitive outcomes; Fagen and Ohr (1990) tested 3-, 7-, and 11-month-olds using the
mobile and train tasks after a 1-week delay and found moderate correlations ($r = .40$ to $.50$) between their performance on these operant tasks as infants and their performance on standardizes cognitive assessments at 2 and 3 years of age.

Operant conditioning methodology maps well to studies conducted with non-human populations, and there is a large body of existing empirical data on memory processing using this approach. Motor sensors may be incorporated to automate data collection with both the mobile and the train tasks. The disadvantage of operant memory tasks is that administration of the measure may take days to allow for training and test periods, although different types of associative learning conditioning protocols (see Fifer et al., 2010) have recently been adopted and have good predictive outcomes as well.

**Visual Recognition Memory.** Early information about infant memory came from Visual Recognition Memory (VRM) studies of looking patterns. The VRM paradigm exploits the fact that infants look more at novel stimuli than at familiar stimuli (Bahrick & Pickens, 1995; Bahrick, Gogate, & Ruiz, 2002; Rose, Feldman, & Jankowski, 2004). In a typical VRM procedure, an infant is familiarized with a visual stimulus (e.g., e.g., Bahrick et al. 1995; 1997; 2002; Morgan & Hayne, 2006). This familiarization period usually occurs for a specified time, or until the infant’s looking time accumulates to a predetermined level or decreases by a predetermined amount. After a delay, the familiar stimulus and a novel stimulus are presented simultaneously and the time that the infant spends looking at each is compared. If the infant looks more at the novel stimulus than at the familiar stimulus (i.e., shows a novelty preference), it is inferred that he or she remembers the familiar one. This task can be used from 6 months of age to 4 years, as well as across the lifespan (see Morgan & Hayne, 2006; Barr, Walker, Gross, & Hayne, 2014). As with the operant conditioning paradigm, parameters of the task and retention
of the task change as a function of age. VRM can also be assessed across modalities (Rose, Gottfried, & Bridger, 1978). For example, Rose and colleagues (1978) gave infants the opportunity to tactually explore an object without seeing it and then gave infants a visual recognition task. Such cross-modal measures provide a measure of memory flexibility. Like the operant conditioning task, VRM has been reported to predict later outcomes (e.g., Rose, Feldman, Jankowski, & Rossem, 2005; Rose et al., 2009). Measures of VRM have been reported to correlate well with other measures of memory, with stability from 2 to 3 years, as well as continuity from infancy through toddlerhood (Rose et al., 2005). Performance on the VRM task has been reported to predict performance on the Bayley Mental Development Index (BMDI), a global measure of cognitive ability administered at 3 years (Rose, Feldman, & Jankowski, 2009), as well as IQ at 11 years (Rose, Feldman, Jankowski, & Van Rossem, 2012; see also Burbacher, & Grant, 2012; Jacobson, 1998; McCall & Carriger, 1993). The VRM task is relatively quick to administer. Measurement using eye tracking has become more common and is recommended as an automated, finer-grained method for data collection (Aslin, 2007). It should be noted that these studies have not gone without criticism regarding which cognitive processes researchers are actually measuring and reporting (e.g., Oakes, 2010).

**Deferred Imitation.** Deferred imitation (DI) is a nonverbal memory paradigm used during early infancy and prior to early childhood starting at 6 months through at least 3 years of age using the age appropriate parameterization. DI provides an optimal measure of memory in preverbal infants because it requires the infant to encode, retain, and retrieve a memory—all without the production of language. In this paradigm, infants (1) see a model demonstrate target actions on an object, (2) internalize the representation of the actions, and then (3) reproduce those actions after a delay (Piaget, 1962). For example, in the widely used puppet imitation task
(Barr, Dowden & Hayne, 1996), a child sees the experimenter remove a mitten from the puppet’s hand, shake it to ring a bell inside, and replace the mitten on the puppet’s hand. This sequence takes 10 seconds and is shown multiple times, with demonstration time varying with the age of the child (e.g., 6 times for 6- and 9-month-olds, 3 times for 12- to 24-month-olds). During the DI test, the infant is given an opportunity to imitate the modelled actions. Overall, baseline performance (i.e., spontaneous production of the target actions) on this task is very low during infancy and has been established across multiple studies. The imitation score is the number of target actions reproduced within a set time limit (e.g., 120s for 6-9 months; 90s for 12-24 months) to reproduce the target actions. To test memory flexibility, infants may also be shown a demonstration of target actions and then tested with a functionally equivalent but perceptually novel object.

Rose and colleagues (2005) tested 1-, 2-, and 3-year-old infants on DI tasks and found that cross-age correlations were highest between 2 and 3 years. DI during the first two years of life also predicts overall cognitive performance at 4 years (Strid, Tius, Smith, Meltzoff, & Heimann, 2006) and memory performance and school readiness at 6 years (Riggins, Cheatham, Stark & Bauer, 2013). The deferred imitation paradigm is very practical for use in infants and toddlers; this is a relatively quick task, only taking a few minutes and attrition rates are low. The task also has predictive validity and discrimination. Performance is correlated across ages and is predictive for cognitive outcomes at age 4 years. However, the range of scores that can be produced from typical DI tasks is limited; researchers may need to consider combining different tasks and potentially creating memory composites with other types of memory tasks in order to increase the variability of outcomes to better predict long-term cognitive outcomes.

**Executive Functions**
Executive functions (EF) are a set of general-purpose control mechanisms that regulate goal-directed behavior (Best & Miller, 2010) and are associated with prefrontal cortex function. The most popular framework for EF (Miyake, et al., 2000; Diamond, 2013) posits three foundational components: updating (the constant monitoring and rapid addition/deletion of information in working memory), inhibition (the purposeful overriding of prepotent responses, also called inhibitory control:) and shifting (the ability to switch between tasks or mental sets, also called cognitive flexibility). The components are correlated with each other to some degree, but do not appear to constitute a completely unitary construct (Best, Miller, & Jones, 2009; Miyake & Friedman, 2012). A common issue in the EF literature is the use of complex tasks that tap into multiple components; for simplicity, researchers typically classify complex tasks by a single EF construct (Miyake et al, 2000). Confusion arises when a task is labeled as a particular EF component by one research team and is then labeled as a different EF component by another research team (e.g., Bell, 2012; Diamond, 2013; Garon, Bryson, & Smith, 2008). Many researchers work around this issue by creating composite measures of EF (e.g., Cuevas, et al., 2014). Similarly, researchers may use the terms of EF, self-regulation, and effortful control interchangeably, leading to debate over underlying components (McClelland & Cameron, 2011). As a result, in the developmental literature, there have been continued calls for conceptual clarity for the constructs of self-regulation, EF, and effortful control (Liew, 2012; McClelland & Cameron, 2012; Zhou, Chen, & Main, 2012).

The development of both the prefrontal cortex and EF are protracted through childhood and early adulthood, with performance on EF tasks exhibiting moderate stability in individual differences by 4 years of age (e.g., Kochanska & Knaack, 2003). Substantial research has revealed that EFs are critical to aspects of optimal development. During early childhood, EFs are
related to school readiness (Bierman, Torres, Domitrovich, Welsh, & Gest, 2009; Blair & Peters, 2003) as well as concurrent and future reading and mathematics performance (e.g., Blair & Razza, 2007; Bull, Espy, & Wiebe, 2008; Clark, Pritchard, & Woodward, 2010; St. Clair-Thompson & Gathercole, 2006). As noted above, performance on EF tasks is linked with prefrontal cortex. This means that development of the prefrontal brain area parallels developmental changes in performance on EF tasks. It is likely that other brain areas are also involved, but EF tasks are typically discussed with respect to frontal functioning and development.

**Inhibitory Control.** Inhibitory control is perhaps the most widely studied EF in young children (Garon et al., 2008; Petersen et al., 2016). Most measurements include a simple response inhibition task, such as the delay of gratification task, which is valid from 24 months of age. Here, children are told they may eat one marshmallow now, or if they wait until the experimenter returns, they may have two marshmallows. This is coded for whether the child waits for the experimenter to return to eat the marshmallow (yes/no) and also the latency for duration until the marshmallow is eaten. Length of delay varies with age of child. The tongue task (Kochanska, Murray, & Harlan, 2000) is valid from 22 months on, and requires children to hold a goldfish cracker on their tongue without chewing it (typically, three trials are used with delays of 10, 20, and 30 s). Performance is the proportion of successful trials. The crayon/gift delay procedure (Calkins, 1997) is detailed in Morasch and Bell (2011) and is also valid from 22 months on. Toddlers are presented with a box of crayons and a blank piece of paper. Before the child touches the crayons, the child is informed that the experimenter needs to leave the room. The toddler is instructed not to touch the crayons, box, or paper until the experimenter returns. The experimenter leaves the room for 60 seconds. Toddlers’ behavior during the delay is scored
with a 0 (colors with crayons), 1 (takes crayons out of box), 2 (picks up box), 3 (touches box), 4 (touches paper), or 5 (does not touch). Latency to touch is also measured.

Complex response inhibition tasks include a variation of the Stroop task (valid from 22 months on), where children are first shown six cards depicting three small and three large fruits and are asked to point to each in turn to ensure they know the names and sizes of the fruits (e.g., “Show me the big apple”). They are then shown three cards each depicting one of the small fruits embedded in one of the larger ones and asked to point to each of the small fruits in turn (e.g., “Show me the small apple”). The score consists of the number (0-3) of small fruits correctly pointed to (Kochanska et al., 2000).

**Working Memory.** Working memory is holding information in mind and mentally working with it or updating it (Diamond, 2013; Garon et al, 2008). Tasks for measuring this construct in toddlers include Spin the Pots/Stationary Pots (valid from 15 months), in which distinct opaque cups are placed upside down and equidistant on a revolving tray called a lazy Susan (Brito, Grenell, & Barr, 2014; Hughes & Ensor, 2005). An object is placed under a cup, and then the child is allowed to find the object, either in a stationary position, or after the cups have been rotated. Scoring on the task includes the number of reaches to retrieve all items and the number of consecutive reaches to same item. Another widely used measure for this construct is the A-not-B/Delayed Response task (from 6 months) and the A-not-B Invisible Displacement task (from 15 months), however, these tasks also require some degree of inhibitory control (Cuevas, Watson, Deater-Deckard, & Bell, 2012; Morasch & Bell, 2011, but see also Thelen et al., 2001 for dynamic systems theory explanations). In the original version of this task (Piaget, 1954), the experimenter hides a toy at location A (it is covered with a cloth) and then allows the infant to search for it. Subsequently, the experimenter then hides the toy at location B. When a
delay is imposed, the infant will typically perseverate and search for the toy at location A. If no
delay is imposed, infants usually search correctly for the toy at location B. For the A-not-B task
the infant’s performance determines the hiding pattern, whereas when measuring the Delayed
Response there is a predetermined hiding pattern (Bell, 2012).

**Cognitive Flexibility.** Cognitive flexibility (a.k.a., set shifting) builds on working
memory and inhibitory control. It involves a changing perspective or approach to a problem or
flexibly adjusting to new demands, rules, or priorities (Carroll, Blakey & Fitzgibbon, 2016;
Diamond, 2013; Garon et al, 2008). The Dimensional Change Card Sort task (DCCS) has been
used as young as 2.5 years of age (Blakey, Visser, & Carroll, 2015) and requires children to sort
cards, usually depicting colored shapes, into trays. The children must sort by one rule first (e.g.,
shape), then they are instructed to sort by a different rule (e.g., color). Most 3-year-olds are able
to sort by the first rule, but once that is changed, often continue to sort by the first rule –
seemingly unable to update the rules of the game and thus make perseveration errors. Four-year-
olds, however, are able to switch rules and sort by the second rule successfully. These age-
related changes in performance on DCCS are typically interpreted as evidence of prefrontal
cortex development. Recent work using Dynamic Field Theory, however, shows that simple
manipulations of the dimensions and features of the card stimuli can improve the performance of
3-year-olds (Buss & Spencer, 2014). There has been recent criticism that cognitive flexibility is
poorly defined and tasks like the DCCS require children to utilize other developing cognitive
skills (e.g., selective attention, working memory, inhibition) and more clearly defined or versatile
tasks tapping into cognitive flexibility are needed (Carroll et al., 2016).

**Other Considerations.** In adults, EF is described as having three latent factors, but
evidence suggests a single factor for young children (e.g., Wiebe, Espy, & Charak, 2008).
Relations between EF tasks seem to change with age (Miller & Marcovitch, 2015), and EF tasks do not inter-correlate before age of 2 or 3 (Diamond, Prevor, Callender, & Druin, 1997). Stability in EF performance is typically seen by age 4 (Jones, Rothbart & Posner, 2003; Kochanska & Knaack, 2003). Current data suggest that infant EF tasks do not correlate with older child EF tasks; however, Rose, Feldman, & Jankowski (2012) suggest that measures of attention, processing speed and recognition memory from 7 and 12 months predict EF at 11 yr. This finding has been replicated using attentional efficiency at 5 months and EF at ages 2, 3, 4 years (Cuevas & Bell, 2014).

**Language**

Language, like other aspects of cognition, is often viewed as composed of many separable components or sub-processes: a phonological system, a dictionary-like lexicon, rules for syntax and grammar, and culturally or contextually shaped aspects of pragmatics. However, the development of these various components has been shown to shape and influence each other (Bates & Goodman, 1997). There is growing appreciation of the dynamic nature of language (Elman, 2004; Christiansen & Chater, 2016; Gogate, Walker-Andrews & Bahrick, 2001); of the role of general cognitive processes in language development (Ibbotson & Tomasello, 2016; McMurray, 2016; Samuelson & McMurray, 2016; Smith, 2013; Gogate & Madhavilatha, 2016); and the role of different environments and cultures in supporting differing developmental trajectories (Hoff, 2006). These points carry two critical implications for early language assessment. First, a more complete picture of a child’s language capabilities may be captured by measurement of the more general processes that support language, rather than by end-state markers of the attainment of, for example, a particular bit of knowledge (e.g., a word, or a grammatical construction). Second, it is critical to measure language in context and as part of a
communicative system that is supported by both the represented knowledge the child has gained and by the people and things in the environment that elicit the conversation.

**Vocabulary.** The MacArthur-Bates Communicative Development Inventory (MCDI) is a standardized caregiver-report vocabulary inventory (Fenson et al., 1994). The MCDI is used widely in the field to measure word learning and language development from 10- to 30-months-of-age. There are two forms, one covering infant development from 10 to 18 months, and another covering toddler development from 16 to 30 months of age. These initial versions of the MCDI have since been expanded to include a third, scaled, form for older children, as well as translations into multiple languages (http://mb-cdi.stanford.edu). The MCDI is widely used with ongoing efforts by multiple researchers to create large, searchable databases of vocabulary data from forms completed by prior studies (http://wordbank.stanford.edu). The advantages are that this measure is validated, reliable, relatively quick, and incorporates a range of words and communicative components from gestures and sound effects to various lexical classes and components of syntax and grammar. Data gained via the MCDI have provided the well-known picture of rapid vocabulary development in the late-infancy and toddler years and the transitions from single words into the two-, and later, three-word utterances that are the beginning of grammar and complex sentence forms. Furthermore, because there are large normative databases associated with the MCDI, it has also been used to examine vocabulary and grammar development in children with a range of skill levels and link language abilities to other aspects of cognitive development (see http://wordbank.stanford.edu/publications, for examples).

These positive points notwithstanding, other factors must be considered with respect to use of the MCDI for cognitive assessment. First, because the MCDI is a caregiver-based checklist, it relies on caregivers’ recall of instances of their child understanding or producing
each of the large number of words on the list. Thus, it may miss words that are used less often or are less likely to be recalled, such as verbs (Sandhofer, Smith, & Luo, 2000). To more accurately estimate the total vocabulary of children based on the MCDI caregiver report, Mayor and Plunkett (2011) developed a corrective algorithm. Second, the criteria used to determine if a child knows a word can differ across caregivers and, indeed, may vary across the same caregiver from time to time (e.g., when the child is younger and older). Third, the MCDI checklist is fixed: it contains the 680 words most likely to be known by a 30-month-old child in the 1990s and thus misses some words that more recently have become common in the vocabularies of more modern toddler cohorts. Fourth, the lengthier and older version of the MCDI can take a very long time to complete as the child’s vocabulary reaches the upper limits of the form, which can also influence caregiver report and extend beyond the caregiver’s ability to recall.

It is also the case that the reliability and validity of the infant version of the form, that measures receptive vocabulary, has been questioned (Feldman, et al., 2000; Tomasello & Mervis, 1994). In response, a number of laboratories have worked to develop performance-based measures of comprehension that can be used with infants in their second year. Friend and Keplinger (2003) report the development of a Computerized Comprehension Task (CCT) that has subsequently been validated in both English- and Spanish-leaning infants (Friend & Keplinger, 2008). These assessments provide a more direct measure of infant knowledge and suggest the potential of efficient, portable tests of early vocabulary development. That said, like the original MCDI, these focus on the end state of the word learning process; quantifying the number of words a child comprehends or produces rather than measuring the process of word learning (but see Hendrickson et al., 2015). In the context of assessment, particularly when following earlier intervention, it may be worthwhile to measure the processing steps that occur
before words are fixed in the lexicon. The Quick Interactive Language Screener (QUILS; Golinkoff, De Villiers, Hirsh-Pasek, Iglesias & Wilson, 2017) is web-based screening tool that includes measures of language process in addition to vocabulary and syntax. Thus, this recently developed tool aims to provide a broader assessment of children’s language ability than vocabulary checklists. The basis of QUILS is a fast-mapping task very similar to referent selection and retention tasks that have received much recent attention in the literature. This recent work builds on an extensive history to research on early word learning in infants and children and has highlighted the complexity of measuring word learning abilities. Detailed review of this literature is beyond the scope of this review, rather we focus on recent developments that point to important considerations for the accurate assessment of word learning abilities.

Reference Selection and Retention Tasks. Reference selection tasks present the child with a visual scene and ask them to indicate a referent using a novel word. Variants of this kind of task in which infants are shown two pictures and a familiar word are played, have been used to show that infants as young as 6- to 9-months of age have some knowledge of common nouns (Bergelson & Swingley, 2012). Other variants ask children to make new mappings “on-line” as the task unfolds. For example, Samuelson and Horst (2008) presented 24-month-old children with two items for which the child’s caregiver indicated they knew a name (e.g., a book and a cat) and one novel item (a unique top) and asked children to “get the blicket.” Children were very good at choosing the novel item (i.e., the top), onto which the novel word (i.e., “blicket”) had been mapped. Retention of this mapping can be assessed by asking children to “find the blicket” after a five-minute coloring break, and in the context of two other novel items to which children had just mapped words.
Samuelson & Horst’s (2008) procedure used 3-Dimensional objects and asked children to make a reaching selection. However, similar tasks using looking or pointing to a visual display have been used with younger children. Differences in the response required of children, not surprisingly, lead to differences in the what is concluded about children’s abilities. In particular, in some looking or pointing tasks 17-month-old infants demonstrate both successful referent selection with novel words and retention of those new name-object mappings (e.g., Halberda, 2003, but see Bion, Borovsky, & Fernald, 2013). In contrast, in reaching versions of the tasks eighteen-month-old toddlers do not perform well on reference selection or retention (Kucker, McMurry & Samuelson, 2018), 24-month-olds perform well on reference selection, but they do not retain the mappings (Samuelson & Horst, 2008), and by 30 months of age children are good at both (see Kucker, McMurray, & Samuelson, 2015a for a review and discussion). Both reference selection and retention, however, can be influenced by a number of external and organismic factors (see e.g., Axelsson & Horst, 2014; Kalashnikova, Escudero & Kidd, 2018; Kucker & Samuelson, 2012; Pomper & Saffran, 2018, Twomey, Ranson, & Horst, 2013). Still other research indicates that children as young as 13 months of age can map and, in some cases, retain novel word-object mappings when only one name and one object are presented at a time (Schafer & Plunkett, 1998; Woodward, Markman, & Fitzsimmons, 1994). However, infant success in these tasks depends on factors ranging from whether mappings are reviewed just prior to a retention test, the length of the delay between mapping and retention probes, and the familiarity of the stimuli (see Horst & Samuelson, 2008 for review). It seems clear that these data are best understood in terms of both situational and developmental processes (Kucker, McMurray, & Samuelson, 2015b), which implies that even small changes in what the child knows can change what they can do the next time they are presented with the task; over longer
timescales, small bits of learning have the opportunity to result in large changes in behavior. Thus, assessment of these abilities must be sensitive to a host of factors including when a given ability is assessed, that can potentially support or hinder the performance of individual children and may do so in different ways for different groups, in particular children at higher risk for developmental delays.

**Novel Noun Generalization.** After the child has mapped and retained a link between a particular object and word, they typically expand the mapping to encompass more instances of the named category. These abilities are often studied with a generalization task using novel nouns. Here, the child is presented with an unfamiliar object that is named. Test objects that match the named exemplar in various properties (shape or material only, shape and color but not material) are then presented and the child is asked which can be called by the name used for the exemplar. Data suggest that, by 24 months of age, children presented with a solid, rigid exemplar will pick test objects that are the same shape as the named exemplar (Landau, Smith, & Jones, 1988). However, children’s generalizations have again been shown to be influenced by a variety of external and organismic factors. Changes in the nature of the response (Samuelson, Schutte, & Horst, 2009), the characteristics of the stimuli, changes in the syntactic frame of the question (Soja, 1992), and interactions with the properties of the objects (Smith, Jones, & Landau, 1996) will affect performance, as will the specifics of an individual child’s vocabulary (Jones, 2003; Perry & Samuelson, 2011). Furthermore, teaching 18-month-old children names for solid things in categories well organized by similarity in shape can create a precocious *shape bias* and lead to a subsequent acceleration in vocabulary development (Samuelson, 2002; Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). Thus, children make decisions about word meaning based on the presented stimuli and their accumulated knowledge, and over time these behaviors
can be changed as small bits of learning accumulate. Behavior is context-bound and knowledge is dynamic; thus, to see what a child knows the investigator should focus on accessing the processes of learning. This is critical for understanding the development of vocabulary and language, but also for the design of new tools to measure that development.

**Other New Tools for Language Measurement.** Our growing understanding of the general processes that shape word learning and language development has enabled new ways of probing what children know. One example is a supportive form test developed by Gordon and McGregor (2014) for 4- to 6-year-olds. This test was inspired by recent work showing that children can use encoded information about what objects were seen where to link novel names to novel referents in ambiguous situations (Samuelson, Smith, Perry, & Spencer, 2011). This task begins in a similar way to the novel noun generalization task. The child is introduced to a novel object that is named. The change comes when the child is later asked to recall that name. During this portion of the task they are shown a sheet of paper with three dots along with the previously named exemplar. The experimenter points to the exemplar and says “Do you remember what this was called?” “Was it the blicket?” (pointing to the first dot), “the blocket?” (pointing to second dot), “or the gazzer?” The child indicates their response by pointing to one of the dots. This task reduces the production demands of recall. Gordon and McGregor (2014) have shown that it can be successfully used with children who are delayed in their vocabulary development, thus providing new insight into atypical developmental outcomes.

Another innovation in assessment of early language development is the use of looking-while-listening and eye-tracking procedures. These are based on a large body of research on adult sentence processing that documents how looking behaviors are coupled to speech processing on a fine-grained timescale (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy,
Examination of infants’ real-time looking behaviors during verbal processing or referent selection tasks has revealed a significant increase in processing efficiency between 15- and 24-months of age, as children learn more words (Fernald et al., 1998; Fernald, Zangl, Portillo, & Marchman, 2008). Furthermore, processing efficiency at 19 months has been shown to mediate the relation between early language input measures and vocabulary size at 24 months in both English- and Spanish-learning infants and to predict measures of language, cognition and working memory at 8 years. Based on this work, a number of laboratories are using eye-tracking to examine the children’s visual exploration and attention during word learning tasks such as referent selection (Roembke & McMurray, 2016; Yu, Zhong, & Fricker, 2012) and novel noun generalization (Lorenz, Mattis, & Samuelson, 2016). These procedures hold the potential for a new view of the underlying decision-making processes as children integrate information presented in the task with their prior knowledge when learning words.

**Socio-Emotional Development**

Emotional and social skills in toddlerhood are key to successful family and peer relationships, academic achievement, and mental health (cf. Beebe, B. & Lachmann, F.M., 2013; Brownell & Kopp, 2007; Eisenberg et al., 1995; 1996; Tronick, 2007). Implicit in such interest in socioemotional competence is the notion that early skills and abilities may forecast a child’s success or failure in the larger worlds of school and peer relationships, and may ultimately be indicators of life success, such as educational attainment, income, and health (Moffitt et al, 2011). At its core, socioemotional competence refers to how successfully a child (1) is able to form and maintain relationships with others, and includes specific skills such as joint attention, affect sharing, attachment, social play and social skills (Burt, Obradovic, Long, & Masten, 2008), and (2) expresses, labels, understands, and manages emotions (Denham, 1998). The next
sections provide a brief review of a representative sampling of socioemotional assessment tools, as well as a conceptual framework for understanding the broader inter-individual and intra-individual contexts in which these skills develop.

**Screening and Diagnostic Tools.** These instruments provide global indicators of a child’s functioning but are less well-suited for measuring specific skill development. The best screening assessments, which are brief, developmentally appropriate, and easy to administer and interpret (Carter, Briggs-Gowan, & David, 2004) include the Ages and Stages Questionnaires: Social Emotional (ASQ-SE; Squires, Bricker & Twombly, 2002) and the Toddler Behavior Screening Inventory (TBSI; Mouton-Simien, McCain, & Kelley, 1997). Both are caregiver-reported assessments, which cover the toddler period (ASQ-SE: 6 to 60 months; TBSI: 12-41 months).

**Comprehensive Socioemotional Assessments.** More comprehensive assessments that include both caregiver-reports and observed behavior are often utilized in research. However, there are relatively few socioemotional caregiver-reported questionnaire measures that demonstrate adequate psychometric properties. Commonly used measures that have been shown to be reliable and valid include the Child Behavior Checklist for 1.5-5 Years (CBCL 1.5-5; Achenbach & Rescorla, 2000), the Infant-Toddler Social and Emotional Assessment (ITSEA; Carter & Briggs-Gowan, 2006a), the Brief Infant Toddler Social Emotional Assessment (BITSEA; Briggs-Gowan & Carter, 2002), and the Toddler Behavior Assessment Questionnaire (TBAQ; Goldsmith, 1996). The CBCL is a 99-item checklist that assesses socioemotional/behavior problems in three domains (internalizing, externalizing, and total problems). The ITSEA is a 166-item, parent-report assessment measure used to identify social-emotional and behavioral problems (internalizing and externalizing) and competencies. The
BITSEA is a shorter version of the ITSEA, with only 42-items that assesses socioemotional behavior problems and specific competencies. The 12 items with the highest loading on each of the ITSEA subscales, as well as 30 items chosen by an expert panel, together comprise the BITSEA. The TBAQ is a 108-item measure that is intended for toddlers aged 16–35 months, and assesses activity level, expression of pleasure, social fearfulness, anger proneness, and interest/persistence. However, it is a tool designed to capture individual differences in emotional expression, rather than skill development in the emotional domain.

**Laboratory Measures of Socioemotional Function.** Laboratory tasks that elicit behaviors of interest may be a more useful tool for assessing socioemotional functioning in infants and toddlers. For example, the Kusche Affective Interview-Revised (Kusche, Greenberg, & Beilke, 1988) consists of a series of open-ended questions assessing metacognitive understanding of emotion and Denham and colleagues (Denham & Couchoud, 1990) developed a puppet task to assess young children’s emotion knowledge. The Laboratory Temperament Assessment Battery (Lab-TAB) is one of the most commonly used behavioral assessments used for research purposes and is comprised of tasks that simulate everyday situations in which individual differences in the expression of emotion, activity level, and regulatory behavior can be observed (Goldsmith & Rothbart, 1999).

A battery of observational tasks may be used to assess young children’s emotional reactivity and regulation strategies during specific kinds of challenges that mimic everyday situations (Calkins & Dedmon, 2000; Calkins, Graziano, Berdan, Keane & Degnan, 2008). Experimenters may elicit positive affect by engaging children in a game of peek-a-boo or by blowing bubbles; fear may be elicited by encouraging children to touch a large realistic moving spider; frustration can be elicited by asking children if they would like a snack, then presenting
the child with a clear plastic container of cookies on the table that cannot be opened. To assess a number of emotion-related skills, these episodes are scored, for example, on the latency, duration, and intensity of particular affective reactions to gauge how responsive the child is under specific conditions of challenge. Global scores of children’s emotion regulation, as well as specific adaptive (i.e. distraction and help-seeking) and maladaptive (i.e., physical and verbal venting) emotion regulation strategies are also scored. The use of specific types of strategies has been found to change with age (Blandon, Calkins, Keane, & O’Brien, 2008) and children with better developed emotion regulation skills are more successful in other domains of functioning such as academic achievement and success with peers (Blandon, Calkins, Keane, Grimm, & O’Brien, 2010; Calkins & Dedmon, 2000; Howse, Calkins, Anastopoulos, Keane, & Shelton, 2003). Significant associations between observed socioemotional skills and young children’s cognitive abilities have been reported (Graziano, Calkins, & Keane, 2011), which suggests that there is some interdependence between the domains of emotion and cognition (Calkins & Bell, 2010), although they may be measured in quite different ways.

**Additional Predictors, Mediators, or Outcomes of Interest**

**Brain Measures**

The following section describes several measures or measurement strategies based on direct assessment of brain activity and brain development.

**Heart Rate Variability.** Precise and timely assessments of neurodevelopmental trajectories and their underlying constructs, including attention, memory and emotion regulation, are significantly enhanced by combining behavioral measures with relatively non-invasive and cost-effective psychophysiological assessments. Specifically, longitudinal measurement of autonomic regulation as assessed by heart rate variability (HRV) at rest, has emerged as an ideal
tool for quantifying brain-behavior development. For example, measures of high frequency heart rate variability, as measured by quantifying respiratory sinus arrhythmia in infants, have been positively correlated with working memory (Marcovitch et al., 2010), shorter visual fixation duration (Richards, 1985), and increased attention in young children (Huffman, Bryan, del Carmen, Pedersen, Doussard-Roosevelt, Porges, 1998). The degree of ANS maturation is linked to many aspects of psychological function and repeated measurement of patterns of heart rate variability at rest could provide a reliable and sensitive marker of the effects of interventions on neurobehavioral trajectories.

**Electroencephalography.** Assessments of power and coherence in electroencephalographic (EEG) activity are efficient and reliable methods for examining developmental changes in brain-behavior relations in infancy and early childhood (Pisani, 2008; Isler et al., 2012, Bell & Cuevas, 2012; Richards, Reynolds, Courage, 2010). EEG provides quantifiable measures of neural activation (i.e., EEG power) and functional cortical connectivity between distinct neural regions (i.e., EEG coherence) and measures the number and strength of synaptic connections, the level of neural connectivity and the degree of maturing brain organization (Marshall, Bar-Haim, & Fox, 2002; Silberstein, Song, Nunez, Park, 2004; Thatcher, North, & Biver, 2005; Grieve, 2010). Neural oscillations as assessed by EEG power and coherence, particularly in frontal regions, are linked to performance (both low and high) on tasks tapping into executive function, cognition, speech and language skills (Brito, Fifer, Myers, Elliott, & Noble, 2016; Gou, Choudhury, & Benasich, 2011; Marshall, Reeb, Fox, Nelson, Zeanah, 2008; Molfese, Morse, & Peters, 1990; Saby & Marshall, 2012; Tierney, Gabard-Durnam, Vogel-Farley, Tager-Flusberg, Nelson, 2012; Williams et al., 2012). Stability of high frequency EEG gamma activity from the newborn to toddler periods has been reported to be
associated with language ability (Gou et al. 2011) and resting frontal EEG gamma power in toddlers is linked to cognitive abilities at 4 and 5 years of age (Benasich, Gou, Choudhury, & Harris, 2008; Gou et al., 2011). Some obstacles to overcome in using EEG with toddlers is that infants and toddlers may not cooperate with putting or keeping the net on their heads and they often move. The development of easy measures of EEG electrode application and ensuring reduced movement artifact are both key to obtaining quality neural activity data.

The Neurodevelopmental MRI Database. There are extreme changes in brain size, synaptogenesis, and myelination in the first year, followed by gradual increases in brain and head size in the second year. The second year signals the beginning of a long series of gradual differentiation of neural growth from about 13 months through puberty; including synaptic differentiation through experience-dependent synaptic pruning and network connectivity through axonal myelination. While there is an empirical literature on neuroimaging and brain measurement during the first year (Fillmore, Richards, Phillips-Meek Cryer, & Stevens, 2015), there is precious little information in this domain during the second year of life. A database that may be used for neuroimaging studies has been constructed (Richards, Sanchez, Phillips-Meek, & Xie, 2015; Richards & Xie, 2015). The Neurodevelopmental MRI Database (https://tinyurl.com/MRIDatabase) provides neuroimaging tools that can be used with infants in the first and second year of life, provides actual data for examination of neurodevelopmental changes in infants and beyond, and is useful for a range of neuroimaging studies (structural MRI, DTI, connectivity). Procedures for structural MRIs in infants from 13 to 24 months need to be developed and should also be accompanied by further enhancement of the average MRI templates in the Neurodevelopmental MRI Database for use in the second year.
Sleep

The role of sleep in infant and toddler neurobehavioral development cannot be overstated. The distinct physiologic sleep states, and the cycle into which they are organized, during both daytime naps and nocturnal periods, are critical for learning, memory, growth, immune functioning, and ultimately for survival (see Gómez & Edgin, 2015; Huber & Born, 2014; Lushington, Pamula, Martin, & Kennedy, 2013). The so-called typical patterns of infant sleep are sometimes over-generalized (for more detail, see Montgomery-Downs, 2008); there are currently no data upon which to base universal sleep recommendations for any age group. For example, sleep need is largely influenced by individual differences and culture (Middlemiss, Yaure, & Huey, 2015). Nonetheless, these factors conspire to make this both a rich and challenging field, which has contributed tremendously to our understanding of early development.

Sleep may be broadly characterized as an upstream regulator of cognitive processing. Polysomnography, the gold-standard for measuring sleep (Iber, Ancoli-Israel, Chesson, & Quan, 2007), is a multi-parametric assessment that includes, at a minimum, electroencephalography (EEG), electrooculography (EOG), and electromyography (EMG). Cardiorespiratory measures needed to identify sleep-disordered breathing include additional sensors to measure air flow, blood oxygenation, snoring, and respiratory effort. Proxy methods for quantifying sleep are also available. Scientific-grade accelerometers, or actigraphs, show varying adequacy for valid and reliable measurement of sleep/wake cycles in infants and toddlers (see Meltzer, Montgomery-Downs, Walsh, & Insana, 2012). However, it is important to note that open-market commercial devices have generally shown poor validity for sleep measurement and care should be taken to ensure that any device used shows strong concurrent validity against the gold standard (for
example, see Meltzer, Hiruma, Avis, Montgomery-Downs, & Valentin, 2015; Kolla, Mansukhani, & Mansukhani, 2016; Roomkham, Lovell, Cheung, & Perrin, 2018). Caregiver reports have also been investigated and reports of new techniques are frequently published (see Spruyt & Gozal, 2011a). Best practices in psychometric translation and validation performance should be required of any emerging subjective assessments (Spruyt & Gozal, 2011b). It is also worth noting that caregivers themselves are susceptible to sleep disturbances, which has a marked influence on episodic memory (Inostroza & Born, 2013), so retrospective reports (or those that caregivers can fill in retrospectively, regardless of instructions) may also be biased.

**Parent-Child Interactions**

The family environment provides important contexts for infant and toddler development. Positive parent-child interactions are critical for optimizing developmental outcomes in the areas of social-emotional and cognitive growth. Vygotsky’s (1978) social constructionist theory proposes that all cognitive functions develop through social interactions, and studies have demonstrated that parent-child interactions do influence the course of cognitive development during infancy and childhood (e.g., Bornstein & Tamis-LeMonda, 2001; Farrant & Reese, 2000; Vygotsky, 1978). A measure of parent-child interactions, as part of the overall assessment, could explain some of the variance in cognitive functioning across participants. The HOME scale (Caldwell & Bradley, 1984), NCAST (Barnard, 1978), Three Bags Task (NICHD ECCRN, 1999), and IGDI-PCI (Carta, Greenwood, Walker, & Buzhardt, 2010) are all parent-child interaction measures that have successfully used in previous large-scale studies of child development (Brito, Ryan, & Barr, 2014).

Communication is a major aspect of parent-child interactions and parental speech directed to young children is crucial for early child cognitive development. Language can help
create emotional bonds between parent and infant, convey knowledge, and promote learning (Papousek, Papousek, & Bornstein, 1985; Thiessen, Hill, & Saffran, 2005). As the child begins to negotiate language, descriptive conversations between the dyad establishes the foundation for complex dyadic interactions and contributes to later cognitive outcomes (Landry et al., 2006). Observational studies of mother-infant interactions have shown that mothers who provide high levels of responsive verbal stimulation during parent-child interactions were more likely to have developmentally advanced infants (Clarke-Stewart, 1973; Bradley & Caldwell, 1977). Hart and Risley (1995) measured language usage during naturalistic interactions in 42 families from diverse demographic backgrounds over a period of two and a half years. Vast differences in words directed towards children (range 500 to 3000 words per hour) predicted child vocabulary size, expressive language and verbal sophistication, and IQ scores at age three (Hart & Risley, 1995; 1999). It is now possible to easily record the amount of language the child is exposed to on a typical day. The Language Environmental Analysis (LENA) is a 2.5-ounce device that fits inside specially designed clothing and continuously records the child’s language environment. The data are uploaded and automatically analyzed by the LENA program software. Frequency of adult words, child vocalizations, and conversational turns, as well as age-based standard score and developmental ages are calculated. These measures are reliable and highly correlated with standardized assessments (Gilkerson & Richards, 2008; Marchman, Martinez, Hurtado, & Fernald, 2017; Xu, Yapanal, & Gray, 2009).

Measuring parent-child interactions within semi-naturalistic laboratory settings or within the home environment may provide vital information regarding the amount of cognitive stimulation the child is exposed to and could help to explain associations between an intervention and cognitive outcomes (e.g., Adamson & Bakeman, 2006; Masur, 1987; Nomikou, Koke, &
The use of “big data” in the bio-behavioral assessments of infants and toddlers using semi-naturalistic measurement is gaining traction within the developmental science field. This approach is best conducted with measurements made on multiple levels: high quality video and audio digitized recordings, telemetric-based psychophysiological measures (e.g., heart rate and galvanic skin response), and mobile eye-tracking. The method presents several technical challenges: measures must be temporally aligned and synchronized during the data acquisition session. Once these challenges are met, analysis of a simple parent-child interaction over two minutes has the potential to answer a variety of questions about a child’s developmental state. Among the practical reasons for a shift to semi-naturalistic measurement are that this “big data” collection strategy allows for use and re-use of the same participants to ask different questions (Adolph, Gilmore, Freeman, Sanderson, & Milman, 2012). Since the entire realm of behavior has been captured and is available in a digital format, it is possible (for example, several years after the initial session) that an experimenter could return to the recording and code some other realm of behavior that was not initially quantified. This big data approach to semi-naturalistic measurement offers a rare opportunity to gather large data sets from toddlers’ groups that to-date have been a challenge for the developmental science community.

**Conclusions**

Although global, standardized instruments for the evaluation of developmental status of infants and toddlers have their place in early assessment, exclusively or predominantly relying on such tests for evaluating the effects of clinical trials or interventions may underestimate or miss specific effects on early cognition. In this review, we put forth plausible candidates for domain-based assessments of neurobehavioral development in toddlers that have been successfully used
across a range of developmental psychology laboratories. Not all cognitive skills were able to be covered within this review. For example, intersensory or crossmodal perception is a vital aspect of early learning (Bahrick & Lickliter, 2000; Bahrick, Lickliter, & Castellanos, 2013) and related to a number of domains mentioned here (Bahrick, 2010; Bahrick, Lickliter, & Flom, 2004; Gogate, Maganti, & Perenyi, 2014). Another limitation of this review is that some domain-based tests are limited, as normative developmental trajectories and psychometric properties (validity, reliability, fidelity, and predictive validity) necessary for making clinical/health claims have not yet been well-established. Despite these limitations, however, the use of such domain-specific assessments may have great potential for increased sensitivity and specificity that may not be evident with more global measures, and for assessing constructs that are most relevant for cross-validation with preclinical animal models and extrapolation to cognitive outcomes later in childhood.

The sensitivity of specific measures of development may represent an advantage in research, but the use of non-standardized measures of behavioral development in studies of early developmental status and interventions present challenges for clinical trials. Regulatory authorities require scientific evidence of a high standard, and clinical studies provide the most robust evidence for regulatory needs. Clinical studies are essential for assessment of efficacy and potential benefit, as well as safety (adverse events). The responsibility rests with industry to adopt best practices for clinical studies, including measurement of specific outcomes (e.g., neurobehavioral assessment); studies intended to contribute to claims of efficacy in cognitive domains need to be designed and conducted with the expectation that they will be subject to independent expert review. Additionally, many of the paradigms and protocols outlined within this review have little or no norms associated with them. If these alternative tasks were to be
included, a more concerted effort among labs to provide validity and reliability data would be needed.

Although there is no direct guidance from governmental regulators as to the criteria for considering the adequacy of outcome measures for clinical trials, the general criteria for the conduct of clinical studies are becoming widely accepted among scientific experts. The European Union (EU) system is the standard to which most of the world adheres, and so studies are typically designed to meet these standards. However, under the EU system, there is flexibility to design the claim around the cognitive domain that is under measurement, as long as the claim is consistent with the outcome measure. The European Food Safety Authority (EFSA) has only limited guidance on the scientific requirements for health-related claims in cognitive development (EFSA, 2012). The guidance refers to the use of “acceptable outcomes,” such as validated neurodevelopmental tests designed to assess the specific domain(s) which is/are subject of the claim and appropriate to the age group being tested. Given the state of the art assessments of cognitive development available in infancy and toddlerhood, the list of acceptable outcomes provided by EFSA for claims in cognitive development seems inadequate, with an overreliance on global standardized assessments.

Assessment of early development requires several conceptual considerations in addition to the obvious practical and empirical issues that researchers typically address. First, development occurs in the context of social and biological processes that influence a toddler’s functioning in complex ways across development. Current perspectives view development as characterized by cascades, whereby growth in one domain affects growth in other domains (Blair et al., 2015). A second important consideration for the assessment of development is that growth does not occur only in a linear, within-domain fashion. For example, it is quite likely that growth
in the domain of emotion is causing growth in the domain of cognition, and vice versa (Calkins & Bell, 2010). Assessments of infant and toddler development would benefit from the inclusion of biological indicators of functioning and measures of the contextual influence on development, as well as a better understanding of how skills in one domain may impact other domains (e.g., attention and executive functioning). We hope that researchers interested in neurocognitive development are encouraged to use one of these domain-specific measures reviewed here that have been reliably used across our own studies as an alternative to or in conjunction with global standardized assessments in future studies with infants and toddlers. This would enable researchers to have alternative options to global outcomes, the capacity to pinpoint specific underlying mechanisms, and further investigate associations between early measures of neurocognitive functioning and later health and educational outcomes.
ACKNOWLEDGEMENTS

This paper was the result of a two-day workshop organized and sponsored by Mead Johnson Nutrition on the state of the art in assessment of neurodevelopment in infants and toddlers. The contents and its organization are solely the work of the authors and do not reflect any direct input from the sponsor.
REFERENCES


Friend & Keplinger, 2008 J *Child Language* 35(1) 77-98


