How Do We Measure Responsive Caregiving? An Evaluation of Video-Coding Instruments to Quantify Parent-Child Interaction

Andrea M. Imhof

University of Oregon

Author Note

Andrea Imhof, Department of Psychology, University of Oregon.

I want to thank Amy Hane, Martha Welch, Ian Murphy, Judy Cameron, Shannon Peake, and Kyndal Howell for their permission and support to implement the video-coding instruments in this study: The Welch Emotional Connection Screen, Simple Interactions, and the Serve and Return Scale. I also want to acknowledge Caitlin Fausey’s contributions to the development of the Conversational Turns measure. Finally, I acknowledge the 24 undergraduate researchers in the Stress Neurobiology and Prevention Lab who cumulatively spent over 4000 hours coding films for this study.

Correspondence concerning this article should be addressed to Andrea Imhof, Department of Psychology, University of Oregon, Eugene, OR, 97403.

Contact: aimhof@uoregon.edu

Abstract

Recent research has identified that responsive caregiving may serve to buffer the effects of early stress exposure for children in high risk settings. In light of this evidence, researchers at the University of Oregon have developed a flexible, strengths-based video-coaching intervention (Filming Interactions to Nurture Development [FIND]) that highlights and reinforces responsive caregiving behaviors in high-risk families. Because the FIND intervention focuses on increasing responsive caregiving, evaluating intervention effects requires methods to objectively quantify changes in parenting behaviors during parent-child interaction. This study evaluates four video-coding instruments designed to quantify specific parenting behaviors that align with the FIND theory of change – maternal sensitivity, emotional connection, and contingent responsiveness. We compare two global coding measures (the Welch Emotional Connection Scale [WECS] and the Serve and Return Scale [SRS]) and two microsocial coding instruments (Simple Interactions and the Conversational Turns [CT] scale), evaluating each across three domains: feasibility, inter-rater reliability, and viability for implementation. Results suggest that while global coding instruments offer significant advantages in their ability to efficiently synthesize interactions, relying on coders’ general impressions comes at the cost of instrument flexibility and inter-rater reliability. In contrast, micro-social coding measures offer highly detailed, reliable data at the cost of increased training and coding time per film. Ultimately, selecting appropriate video-coding instruments to evaluate FIND requires balancing evaluation criteria across several domains. We discuss the viability of these instruments across diverse contexts and settings, their alignment with the FIND theory of change, coding and training time required, and future considerations relevant to the selection of coding instruments to assess intervention effects.

*Keywords:* responsive caregiving, video-coding, maternal sensitivity, contingency

How Do We Measure Responsive Caregiving? An Evaluation of Video-Coding Instruments to Quantify Parent-Child Interaction

**Introduction**

**Spotlight on Responsive Parenting**

Early environmental experiences have a critical impact on brain development: a growing body of evidence points to early childhood (0-3 years) as a critical period in the majority of cognitive and socio-emotional domains (Cassidy & Shaver, 2018; Shonkoff & Phillips, 2000; Stiles & Jernigan, 2010). Within this developmental window, children’s interactions with primary caregivers play a key role in shaping the quality of the early environment. Responsive parenting has been associated with several aspects of cognitive and socio-emotional development (Bornstein & Tamis-LeMonda, 1989; Landry, Smith, Swank, Assel, & Vellet, 2001; Landry, Smith, & Swank 2003, 2006), including language (Adamson, Kaiser, Tamis-LeMonda, Owen, & Dimitrova, 2018; Hirsh-Pasek & Burchinal, 2006; Tamis-LeMonda, Bornstein, & Baumwell, 2001; Tamis-LeMonda, Kuchirko, & Song, 2014), executive function (Bernier, Carlson, & Whipple, 2010; Blair, Granger, Willoughby, Mills-Koonce, & Cox, 2011), and emotion regulation (Mathis & Bierman, 2015; Morris, Criss, Silk, & Houltberg, 2017). In addition, parent sensitivity and warmth have been linked with attachment security and positive socio-emotional development (Davidov & Grusec, 2006; Deans, 2018; De Wolff & van IJzendoorn, 1997; Frick et al., 2017; Raby, Roisman, Fraley, & Simpson, 2015; Zhou et al., 2002). Parent contingent responsiveness has further been associated with language development, early literacy skills, and growth in other cognitive domains (Bornstein, Putnick, Cote, Haynes, & Suwalsky, 2015; Farah et al., 2008; Golinkoff, Can, Soderstrom, & Hirsh-Pasek, 2015; Holler, Kendrick, Casillas, & Levinson, 2016; Hubbs-Tait, Culp, Culp, & Miller, 2002; Tamis-LeMonda et al., 2001).

In addition to the wealth of evidence supporting the link between responsive parenting, cognitive development, and socio-emotional skills, there is evidence that the opposite may also be true: disruptions to responsive care can have significant physiological, neurobiological, and socio-emotional consequences for infants and children (Bruce, Fisher, Pears, & Levin, 2009; Egeland, Sroufe, & Erickson,1983; Pollak, Cicchetti, Hornung, & Reed, 2000). Evidence from animal models clarifies how maternal interactions impact physiological and psychological stress responses in offspring. For example, Ivy, Brunson, Sandman, and Baram (2008) found that environmental stress (i.e. restricted access to nesting materials) changed the way rat mothers interacted with their pups, causing increased neglect and disruption to normal nurturing interaction patterns. These fragmented maternal interactions provoked chronic stress in rat pups, triggering profound changes in the offspring’s HPA axis regulation. Ivy and colleagues’ experimental manipulation suggests that maternal responsiveness may serve as a critical mechanism underlying the (dys)regulation of early-life stress experiences. Evidence from research with non-human primates corroborates this theory, highlighting how separation, neglect, and disruptions to typical responsive caregiving behaviors can have devastating impacts on offspring (Harlow & Suomi, 1971; Nelson & Winslow, 2009; Sanchez et al., 2005; Sanchez, 2006; Zhang, 2017;). From models of separation and neglect (e.g. Clarke, Kraemer, & Kupfer, 1998; Harlow, Dodsworth, & Harlow, 1965; Sanchez et al., 2005) to disruptions in maternal caregiving behaviors (Maestripieri, McCormack, Lindell, Higley, & Sanchez, 2006; Sanchez, 2006; Szyf, Weaver, & Meaney, 2007), these animals models provide vivid examples of how disruptions to responsive caregiving can impact socio-emotional, neurobiological, and cognitive development in offspring.

For practical and ethical reasons, researchers are unable to conduct these randomized controlled exposures to early stressful experiences with humans. As a result, it is much more challenging to uncover these critical mechanisms in the context of human caregiving relationships. Despite the absence of random assignment and experimental manipulation, researchers have used evidence from institutional care to highlight the profound impact of neglect and disrupted parenting on children in orphanages (De Bellis, 2005). Specifically, research has found that children who experience severe neglect in institutional settings show abnormalities in brain structure (Eluvathingal et al., 2006; Hanson et al., 2013; Sheridan, Fox, Zeanah, McLaughlin, & Nelson, 2012) and diminished electrical activity in the brain measured through electroencephalography (EEG) (Marshall, Fox, & the BEIP Core Group, 2004; Parker, Nelson, & the BEIP Core Group, 2005a, 2005b; Tarullo, Garvin, & Gunnar, 2011). Neglect is also associated with abnormal activity patterns in the prefrontal cortex, the amygdala, and the hippocampus – regions associated with executive function, emotion, attention, and stress regulation (Edmiston et al., 2011; Maheu et al., 2010; Mehta et al., 2009; Tottenham et al., 2010). Children who have experienced neglect also show significant HPA axis dysregulation, marked by altered cortisol patterns throughout the day (Bruce et al., 2009; Dozier, Manni, et al., 2006; Fries, Shirtcliff, & Pollak, 2008; Gunnar, Morison, Chisholm & Schuder, 2001; Koss, Hostinar, Donzella, & Gunnar, 2014). Such children also have higher rates of disrupted (i.e. disorganized, insecure) attachment behaviors (Carlson, Cicchetti, Barnett, & Braunwald, 1989; Egeland & Sroufe, 1981; Hesse & Main, 2000), higher rates of emotional and behavioral problems (Bakermans-Kranenburg et al., 2011), and increased difficulties with attention and executive function (Bos, Fox, Zeanah, & Nelson, 2009; Kreppner, O’Connor, Rutter, & English Romanian Adoptees Study Team, 2001; Pears, Bruce, Fisher, & Kim, 2010), which impact later academic outcomes (Loman, Wiik, Frenn, Pollak, & Gunnar, 2009).

While these models of early disrupted caregiving imply diminished potential for children exposed to high levels of stress, recent research suggests that parent responsiveness may serve to buffer the effects of early stress exposure (Blaisdell, Imhof, & Fisher, 2019; Fisher, Beauchamp et al., 2016; Flannery, Beauchamp, & Fisher, 2017; Gunnar, 1998). In longitudinal studies of institutionalized children, researchers found that children who were adopted before the age of 3 show increased socio-emotional and cognitive functioning compared with children who remained in institutions (Bick & Nelson, 2016; Fox, Almas, Degnan, Nelson, & Zeanah, 2011). These studies suggest that increases in responsive caregiving during the first few years can reverse negative trajectories for children exposed to severe neglect. Further evidence from foster care interventions shows that systematic changes to responsive caregiving can reverse the physiological symptoms of toxic stress, improving cognitive function and normalizing electrical brain activity and diurnal cortisol patterns previously disrupted by early stress exposure (Bernard, Dozier, et al., 2015; Bernard, Hostinar, et al., 2015; Bruce et al., 2009; Dozier, Peloso, et al., 2006; Dozier, Peloso, Lewis, Laurenceau, & Levine, 2008; Fisher & Kim, 2007; Fisher, Stoolmiller, Gunnar, & Burraston, 2007; Nelson et al., 2007). This buffering effect has been replicated in other high-risk populations outside of the foster care system; for example, Kim-Cohen, Moffit, Caspi, and Taylor (2004) found that maternal warmth contributes to child resilience in the face of socio-economic deprivation. In a longitudinal sample, Farrell, Simpson, Carlson, Englund, & Sung (2017) found that maternal sensitivity may help to buffer the effects of early stress exposure on adult health outcomes. Recent research also highlights the role of back and forth interactions as a buffer against poor language outcomes in high risk families (Ludy-Dobson & Perry, 2010; Vernon-Feagans & Bratsch-Hines, 2013). Vernon-Feagan & Bratsch-Hines’ (2013) research further extended the responsive caregiving hypothesis to show how responsive interactions in childcare can buffer against stressful experiences and risk factors at home. Overall, the generalization of this buffering effect across contexts suggests that responsive caregiving may be an important target for interventions designed to reverse the effects of early stress exposure (Bruce, Gunnar, Pears, & Fisher, 2013; Farrell et al., 2017).

**Targeting Responsive Caregiving: The FIND Intervention**

Rooted in evidence that responsive parenting may be a critical mechanism driving positive child outcomes in the face of early adversity, researchers at the University of Oregon have developed a scalable video-based intervention to identify and increase responsive parenting behaviors in high-risk families. Filming Interactions to Nurture Development (FIND) is similar to prior interventions (e.g. Video-Feedback Intervention to Promote Positive Parenting [VIPP; Juffer, Bakermans-Kranenburg, & IJzendoorn, 2008] & the Attachment and Biobehavioral Catch-Up Program [ABC; Dozier, Peloso et al., 2006]) that utilize a video-coaching framework to teach caregivers supportive and responsive parenting strategies. The FIND Intervention frames responsive caregiving using the metaphor of “serve” and “return”, helping parents to conceptualize contingent responsive interactions using accessible language (Fisher, Frenkel, Noll, Berry, & Yockelson, 2016; Shonkoff & Bales, 2011). Building on research that highlights the importance of micro-social interactions (Fisher & Gilliam, 2012; Patterson & Reid, 1984) FIND coaches utilize edited 3-5 second video clips to highlight positive examples of caregivers interacting with their children. The intervention targets 5 primary examples of contingent responsive interactions through the Find 5 Elements: a) Sharing the Focus b) Supporting and Encouraging c) Naming d) Back and Forth and e) Endings and Beginnings. In lieu of traditional didactic approaches, FIND uses individualized video-feedback to provide positive examples of caregivers using the FIND 5 elements during play. Strengths-based approaches such as those employed in FIND (e.g. Kemp, Marcenko, Lyons, & Kruzich, 2014), have been successful in improving caregivers’ self-efficacy and reducing parent stress (Schindler, Fisher, & Shonkoff, 2017).

Preliminary evaluations of FIND show promising results that this intervention model may lead to significant improvements in parent outcomes. In an early pilot of FIND with 32 families, Guiliani, Beauchamp, Noll, & Fisher (2019) found that FIND increased mothers’ endorsement of child-supportive parenting traits in the parenting self-evaluation task (PSET). Guiliani and colleagues also found functional brain changes during inhibitory control tasks in mothers who participated in the FIND intervention compared with controls. These findings constitute one of the first published evaluations of how the FIND intervention impacts parent outcomes, leaving us with promising pilot evidence that FIND may lead to positive brain and behavior changes for caregivers.

While these preliminary findings support positive changes in parent outcomes following the FIND intervention, we have yet to systematically evaluate how this intervention changes parent-child interaction. Given that the intervention specifically targets improving parenting behaviors, measuring these interactions has become a primary goal in evaluating the efficacy of FIND. The present study focuses on evaluating several video-coding instruments designed to measure responsive caregiving behaviors. Our primary aim is to identify the most promising behavioral coding tools for future use in evaluating the impact of FIND on responsive caregiving behaviors.

**Measuring Parent-Child Interaction**

**Defining responsive care: maternal sensitivity vs. contingent responsiveness.**

The question of how best to measure responsive caregiving has a long history, focusing on a number of caregiver skills that include recognizing and providing support for the child’s needs, providing positive reinforcement, engaging in appropriate limit setting and non-harsh discipline, and encouraging extended back and forth interactions (Pettit, Bates, & Dodge, 2007; Raby, Lawler et al., 2015). Within several studies of parent-child interactions with young children, two primary dimensions of parenting have emerged: maternal warmth/sensitivity (e.g. Davidov & Grusec, 2006; Leerkes, Blankson, & O’Brien, 2009; Pederson et al., 1990) and contingency (e.g. Keller, Lohaus, Volker, Cappenberg, & Chasiotis, 1999; Keller Lohaus, Volker, Elben, & Ball, 2003; Kuchirko, 2017; Tamis-LeMonda et al. 2014).

Maternal sensitivity, originally described in Ainsworth, Blehar, Waters, & Wall’s 1978 seminal research on infant attachment, includes positive tone, nurturance, empathy, and outward affection. Lohaus, Keller, Ball, Elben, & Volker (2001) linked Ainsworth’s original construct with maternal warmth, creating a combined construct that encompasses aspects of emotional connectedness and empathetic responsiveness to child behaviors. Caregiver sensitivity and warmth have been most consistently linked to the quality of the caregiver-infant attachment, which predicts positive mental health and socio-emotional development (Bigelow et al., 2010; Bretherton, 2013; McElwain & Booth-LaForce, 2006; Stams, Juffer, & van IJzendoorn, 2002). Although there is still some variation in how caregiver warmth and sensitivity are measured, behavioral coding techniques grounded in this theoretical perspective tend to highlight emotional connectedness, synchrony, and caregiver affect in response to child cues (Bohr, Putnick, Lee, & Bornstein, 2018; De Wolff & van IJzendoorn, 1997; Meins, Fernyhough, Fradley, & Tuckey, 2001; Tryphonopoulos, Letourneau, & DiTommaso, 2016). For example, the Ainsworth Maternal Sensitivity Scales (AMSS; Ainsworth, 1969), measure the caregiver’s ability to notice, interpret, and respond appropriately to an infant’s signals, and are highly correlated with attachment security in the strange situation task (Pederson, Bailey, Tarabulsy, Bento, & Moran, 2014). In comparison, the Emotional Availability Scales (EAS; Biringen, Robinson, & Emde, 1998) integrate Ainsworth’s original definition of maternal sensitivity with emotional availability, including caregiver emotion signaling within its conceptualization of sensitivity and responsiveness (Bohr et al., 2018; Tryphonopoulos et al., 2016). Overall, behavioral coding instruments that fall within this dimension tend to focus on caregiver *sensitivity*, *emotional affect, and synchrony* within the context of the larger interaction.

In contrast, contingency/predictability represents a separate dimension of caregiver responsiveness, characterized by the quality and quantity of back-and-forth interactions between parents and their children. Grounded in social learning theory (Bandura, 2001), contingent social responsiveness conceptualizes caregiver-child interactions as mutually dependent; instead of primarily focusing on parent actions, this model focuses equally on reciprocal contributions from the infant and caregiver (Dunst, Loww, & Bartholomew, 1990). Contingent social interactions are a critical precursor to language development, early literacy skills, and growth in other cognitive domains (Dunst et al., 1990; Dunst & Kassow, 2008; Bornstein & Tamis-LeMonda, 1989, 1997; Kochanska, Forman, & Coy, 1999; Tamis-LeMonda et al., 2001). For example, verbal back and forth interactions have long been associated with infant acquisition of communicative competence (Bornstein et al., 2015; Demaio, 1982; Dunst et al., 1990; Golinkoff et al., 2015), a construct that has led to the popularization of research on “conversational turns” in the literature today (Gilkerson et al., 2017; Roberts & Kaiser, 2011; Romeo et al. 2018; Tamis LeMonda et al., 2001).

Bornstein, Tamis-LeMonda, Hahn, & Haynes (2008) further dissected aspects of contingent responsiveness by using a moment-to-moment event-coding scheme to record several different categories of parent responses during play (e.g. affirmation, imitations, questions etc.). They found that specific domains of responsiveness (e.g. questions) that focus on eliciting input from the child directly shape children’s communicative skills, amplifying connections between language its meaning. Bornstein et al.’s detailed breakdown of maternal responsive behaviors further elucidates the separation between maternal sensitivity and contingent responsiveness first proposed by Lohaus et al. in 2001; this theoretical split remains relevant to how researchers conceptualize and measure responsive caregiving behaviors today.

**Level of zoom: global vs. microsocial coding.**

Beyond considering *which* behaviors best embody “responsive caregiving,” there are also differing perspectives on *how* to best quantify these behaviors. While parent self-report measures can offer some perspective on parent/child interactions outside of a laboratory setting (e.g. Davidov & Grusec, 2006), self-reported responses to questionnaires are often biased by caregivers’ tendencies to respond in a socially desirable manner (Eddy, Dishion, & Stoolmiller, 1998; Gardner, 2000). Therefore, observational measures (e.g. live behavioral coding, video-coding) are considered to be the gold standard for quantifying parent-child interactions (Bardack, Herbergs, & Obradovic, 2017). In recent decades, several dozen different observational coding schemes have emerged, triggering a separate field of research dedicated to describing different frameworks and approaches for quantifying the frequency, duration, and quality of different behaviors within social interactions (Alexander, Newell, Robbins, & Turner, 1995; Floyd, 1989; Hirschmann, Diemann, Schmelzer, & Pietschnig, 2018; Margolin et al., 1998). These frameworks have helped to differentiate two distinct dimensions of measurement: *microsocial* vs. *global* coding schemes (Hawes, Dadds, & Pasalich, 2013; Markman, Leber, Cordova, & St. Peters, 1995).

In global systems, coders use guidelines to summarize interactions based on general impressions of specific dimensions (e.g. disengagement, encouragement, shared focus). These coding schemes tend to require high levels of inference, asking raters to provide their overall impression of a specific behavior across an entire observational period. Many research groups have favored global coding systems over the years because they are less expensive and require significantly fewer coding resources than microsocial coding systems (Bardack et al., 2017; Hawes et al., 2013; Markman et al., 1995). In the past two decades, several global coding schemes have emerged, increasing the feasibility of drawing connections between parenting behaviors and child outcomes (e.g. Caspi et al., 2004; Eisenberg et al., 2005; Roggman, Cook, Innocenti, Norman, & Christiansen, 2013; Zhou et al., 2002).

However, despite the popularity of global rating systems, this method of coding is susceptible to systematic bias: when coders summarize their impressions of behaviors across several minutes of interaction, they inherently rely on subjective inferences to quantify complex interactions. Coders can be influenced by prominent events within the observation period or can be overly influenced by initial or final impressions of the caregiver and child, impacting reliability (Bardack et al., 2017; Gardner, 2000). Further, because global rating systems tend to summarize behavioral impressions, they do not retain the granularity of microsocial coding schemes, losing specific details about the sequence and frequency of specific events (Hawes et al., 2013). Despite these limitations, global coding schemes present a promising option for cost-effective behavioral coding that can be applied across several contexts of parent-child interaction.

Microsocial coding schemes, in contrast, quantify behaviors at a moment-to-moment level, capturing the frequency, sequence, and duration of specific events as they occur during an observation period. Popularized in the 1960s and 70s by Patterson and colleagues at the Oregon Social Learning Center (Patterson, Reid & Dishion, 1992; Patterson & Reid, 1984; Dishion, Gardner, Patterson, Reid, & Thibodeaux, 1983), microsocial coding allows researchers to pinpoint specific behaviors within parent-child interactions that predict child and family outcomes. By homing in on specific behaviors at the moment-to-moment level, Patterson et al. (1982) were able to identify the specific behavioral patterns underlying healthy and dysfunctional relationships, forming the basis for more targeted interventions (Eyberg, Nelson, & Boggs, 2008; Hawes et al., 2013; Markman & Notarius, 1987).

Beyond their ability to highlight the importance of specific behavioral patterns within dyadic interaction, microsocial coding techniques also eliminate many of the concerns about systematic bias and subjectivity that apply to global codes. Typically characterized by detailed definitions of mutually exclusive behaviors, microsocial coding instruments require coders to follow complex rules that eliminate most subjectivity and inference (Alexander et al., 1995; Dishion, Mun, Tein, Kim, & Shaw, 2017). These coding schemes can vary in their methodology, requiring raters to count the frequency of specific behaviors (e.g. Aspland & Gardner, 2003; Patterson, 1982), or using short-interval ratings to quantify contingent parent and child behaviors (e.g. Putnam, Spritz, & Stifter, 2002). Both types of microsocial coding (short-interval vs. frequency-based) offer significant advantages over the level of detail provided by global ratings, but have significant drawbacks in their labor cost. Depending on the complexity of the coding scheme, it can take several hours to code a short (5-10 minute) segment of parent-child interaction (Hawes et al., 2013; Aspland & Gardner, 2003).

**Limitations of current observational coding schemes.**

Beyond the lack of consensus about *what* constitutes responsive caregiving (sensitivity vs. contingency) and *how* to code it (microsocial vs. global), a recent study by Bohr et al. (2018) identified another challenge to measuring behavioral interactions: many of the video-coding instruments that claim to measure the same construct may actually be measuring different things. Specifically, Bohr and colleagues compared four well-known video-coding instruments that claim to measure maternal sensitivity and emotional connection: the Ainsworth Maternal Sensitivity Scales (AMSS), the Emotional Availability Scales (EAS), the NCAFS Feeding Scale, and the Mini MBQS-VR. Of the four codes measured, 10-91% of the variance between scales was unshared, indicating that these coding tools may not be similar enough to be used interchangeably. Bohr et al.’s findings expose a clear gap in the field, inviting better measurement tools and evaluation standards for these coding instruments.

Another limitation of existing measures is that most of them require laboratory -based, high quality films and highly trained, well-resourced research teams to conduct the ratings. While these rigorous standards (e.g. Eyberg et al., 2004; Brady-Smith, O’Brien, Berlin, Ware. & Brooks-Gunn, 1999) are seen as necessary to eliminate bias and measurement error in scientific research, they also increase costs and challenge the feasibility of the evaluation process. Thus, measures need to be both high quality and feasible in real-world settings.

**Aims of this Study**

In light of the controversies surrounding many of the older video-coding measures (e.g. Bohr et al., 2018; Lotzin et al., 2015), several new coding instruments have emerged that aim to address the weaknesses of their predecessors. This study seeks to evaluate four of these newer video-coding instruments, establishing their feasibility and utility as measures of responsive caregiving. In the present study, we evaluated the feasibility, reliability, and implementation viability of these four coding measures, with the goal of identifying better tools to evaluate change in parenting behaviors following intervention. As a part of this project, we chose measures that represented a mixture of the two primary theoretical perspectives of responsive caregiving: maternal sensitivity and contingent responsiveness. We also selected a variety of moment-to-moment measures and global rating scales to investigate a broad spectrum of measures that might be sensitive to change and also scalable for evaluation in different contexts.

Given that the majority of existing coding systems are extremely resource intensive and limited in scope, a secondary aim of this study is to evaluate the feasibility and flexibility of different video-coding tools. We tested a diversity of different approaches to behavioral coding systems, considering the level of analysis (global ratings vs. microsocial coding), data quality, and training time as a part of our evaluation process. This novel approach -- coding the same videos using a broad range of coding tools -- allows us to directly compare different video-coding methodologies to identify instruments with the most potential to evaluate FIND intervention effects across a variety of settings.

Finally, one of the unique aims of the FIND intervention is to conduct ongoing impact evaluations at scale: that is, investigations of effectiveness across a wide variety of community settings to maximize impact for families at risk (Fisher, Frenkel, et el., 2016). Given the wide range of settings where FIND is currently being implemented – from pediatric primary care settings to homeless shelters – it is critical that our evaluation tools mirror the flexibility of the intervention itself. Accordingly, an additional aim of this study is to identify flexible video-coding instruments that can be implemented reliably across several different contexts, maximizing the feasibility of using this tool in community-based research.

**Methods**

**Video-Coding Instruments**

The present study evaluates four video-coding instruments designed to capture different dimensions of parent/child interaction and parent responsiveness. The measures selected comprise a diverse set of theoretical perspectives (maternal sensitivity, emotional connection, and contingent responsiveness) and coding methods (global and microsocial), and were chosen because of their alignment with the FIND intervention theory of change (Fisher, Frenkel, et al., 2016). Coding manuals for each measure can be found in Appendices A - G.

**Welch Emotional Connectivity Scale**

The Welch Emotional Connectivity Scale (WECS) is a global measure of parent sensitivity and emotional connectivity between mothers and their children. The WECS has been validated as a screener for rating mother-infant emotional connection (Hane et al., 2019), and is used as an efficient video-scoring measure to characterize maternal/infant relationships that predict longitudinal child outcomes (Fagan et al., 2019; Frosch et al., 2019). Mother/child dyads are filmed in a 10-minute face-to-face interaction, with the infant seated in a booster chair or on the parent’s lap. Video coders score four behavioral dimensions of emotional connection:

*a) Attraction to Each Other*

*b) Affective Vocal Communication*

*c) Affective Facial Expressiveness*

*d) Sensitivity/Reciprocity to Each Other*

Each item is rated on a 3-point scale from low (=1) to high (=3), primarily focusing on whether that specific dimension was involved with *establishing or maintaining a connection within the dyad* (Hane et al., 2019). After scoring each of the 4 primary dimensions, coders make a final judgement about whether the dyad was emotionally connected (yes/no). (See Appendix A).

**Serve and Return Scale**

The Serve and Return Scale (SRS) is a 16-item global coding system developed by researchers at the University of Oregon to directly assess responsive parenting (Fisher, Berry, Howell, & Noll, 2016). This measure is grounded in the core FIND concepts of “serve” and “return” (Fisher, Frenkel, et al., 2016; Shonkoff & Bales, 2011), evaluating the quality and quantity of serve and return interactions between caregivers and their children during free-play. To score using the SRS, coders watch a film once, taking notes and tallying behaviors that align with specific items on the scale. These tallies and notes are then used to inform global ratings of how frequently specific behaviors occurred during the film. The SRS consists of 3 subscales:

*b) Supporting and Encouraging*

*c) Naming*

*e) Caregiver Initiated Activities*

Each of these subscales is further broken down into individual items, which are scored on a scale of 1 (“rarely”) to 3 (“often”) based on how frequently these behaviors occurred during the film. Coders can score a fourth option (“No Opportunity”) if that specific behavior was not observed during the film.

The SRS scale is not yet published. The scale was originally developed by the FIND consulting team in tandem with the development of the intervention, with the goal of directly evaluating each of the FIND 5 Elements. Previous attempts to validate this scale yielded concerns about the reliability and psychometric properties of specific sub-sections of the original scale; as such, video-coding teams for this study used a modified version of the original SRS that included 3 out of the 6 subscales designed for the original SRS measure. (See Appendices B and C).

**Simple Interactions**

Simple Interactions (SI) is a moment-to-moment video-coding instrument developed to quantify three primary dimensions of dyadic interaction: Connection, Reciprocity, and Progression (Murphy et al., 2018). Adapted from a global rating system that has been implemented in child care centers, classrooms, and orphanages (Li & Julian, 2012; Akiva, Li, Martin, Homer, & McNamara, 2017; Li, 2014), Simple Interactions considers the quality and quantity of interactions between caregivers and children on a moment-to-moment basis using Noldus Observer XT video-coding software. The overall instrument uses one primary glossary of terms (see Appendix D), which outlines specific definitions for core concepts and critical behaviors (e.g. “serve”, “return”, “interaction in joint activity”). Each subscale is scored using a flow-chart approach to guide coder decisions, ultimately leading to a 3-point rating system across each dimension from X (“low”) to Z (“high”). For the purposes of this study, we selected the first two subscales -- Connection and Reciprocity -- for evaluation, because they align most closely with our project aims.

**Connection***.*

The SI Connection scale is a measure of synchrony between parent and child, quantifying the level of emotional connection and joint activity. Connection is scored on a moment-to-moment basis using a flowchart of decisions that evaluates synchrony of affect, engagement in joint activities, physical proximity, and similar looking/talking/touching behaviors (See Appendix E). Using this flowchart, each frame can be assigned one of four different codes: CX (low), CY (medium), CZ (high), or No Code. After coding, the uncoded (“No Code”) segments are subtracted out from the rest of the coded film, and each code is assigned a specific weight:

*%CX\*100 + %CY\*300 + %CZ\*500 = Final Score*

Connection scores for an individual range from 100 (low emotional connection between caregiver and child) to 500 (high emotional connection).

**Reciprocity***.*

The SI Reciprocity scale is a micro-social measure of serve and return, allowing us to quantify serve and return exchanges over the course of a film. Reciprocity is scored across 15-second intervals, with each interval assigned one of three codes based on the quality and quantity of interaction within that 15-second clip: RX (low), RY (medium), RZ (high), or No Code. Coders evaluate dyadic engagement, interaction in joint activity, leadership, and the balance of serve-and-return exchanges, with balanced reciprocal interactions earning the highest (RZ) scores. One-sided exchanges (e.g. primarily caregiver serves) are scored lower on the Reciprocity scale (see Appendix F). After coding, each Reciprocity code is assigned a specific weight to calculate a final Reciprocity score between 100 and 500:

*%RX\*100 + %RY\*300 + %RZ\*500 = Final Score*

**Conversational Turns (CT)**

The Conversational Turns Scale (CT) is a simple coding scheme designed to record timing, quantity, and length of caregiver and child utterances. The scale is designed for use in ELAN Linguistic Annotator, a software system designed for processing audio and video files, and can be used to code dyadic interaction across all ages and languages. The CT scale was adapted by the author (A. Imhof) and C. Fausey from similar coding systems for annotating utterances that are well-established in the literature (e.g. Pretzer, Lopez, Walle, & Warlaumont, 2019). The CT Scale is scored across 3 tiers: 1) Parent Utterances 2) Child Utterances and 3) Coded Segment. The first two tiers record specific timing and segment length for caregiver/child vocalizations, and the third “Coded Segment” tier records the specific start and stop times for the segment of film that was coded. Utterances are defined and coded according to specific criteria outlined in Appendix G.Overall, this scale provides an estimate of the specific timing and length of caregiver and child utterances during filmed interaction.

**Coder Training**

A total of 24 undergraduate research assistants were recruited and trained by the lead author to serve as video-coders for this study. Coders were divided into teams of 3-4 and trained on one of the four coding scales. Each team was trained using protocols specific to each coding instrument; training involved group didactics, individual practice coding assignments, and group meetings for discussion and conferencing. Training films contained 5-10 minutes of dyadic free-play interactions in both lab and home-based settings; these films were taken from previous FIND implementations and primarily featured dyadic interactions with infants aged 6-36 months. These training films were supplemented by gold standard films designed to accompany each coding instrument, when possible. Teams met weekly throughout the training phase to review reliability statistics, compare scores, and troubleshoot discrepancies from gold standard films created by expert coders. Individual coders were expected to reach a minimum reliability threshold with Gold Standard films to complete the training process (see “Reliability” for more information about how reliability was calculated for each instrument). Training and coding took place over the course of 10 months.

**Data for the Current Study**

This project utilized films from a 5-year randomized control trial of the FIND intervention in Denver, Colorado (“ACF”). This dataset contained 176 films of caregiver-child interaction, each containing 6 minutes of free-play recorded in the family’s home. Caregivers and children in this dataset were recruited from Early Head Start programs in the Denver area and were screened for evidence of toxic stress to qualify for inclusion in the intervention trial (e.g. abnormal diurnal cortisol patterns, endorsement of 3+ Adverse Child Experiences, qualification for public welfare resources); therefore, this dataset represents a high-risk sample of caregivers and children. The majority of families included in this sample were bilingual: 69% (n=122) films contained interactions in Spanish, 29% (n=51) were recorded in English, and 7% (n=12) contained interactions in both English and Spanish. Of the films coded for this project, the majority (n=109) were gathered pre-intervention, representing baseline caregiver/child interaction. The remainder (n=67) were gathered post-treatment. Of the 109 families included in this study, two thirds (n=73) were assigned to receive the FIND intervention, while the remainder (n=36) served as a control condition, receiving Early Head Start services only, without FIND coaching. All coders were blind to condition assignment and pre/post film status throughout the coding process; we report feasibility, reliability, and implementation considerations across all 176 coded films.

**Dimensions of Evaluation**

After selecting these four measures (WECS, SRS, SI, and CT), we conducted a systematic evaluation of each instrument across three primary domains: a) feasibility b) reliability and c) implementation viability:

**Feasibility.**

Given that the goal of this research was to identify video-coding instruments that could be used to evaluate the FIND intervention, it was critical that we considered whether our chosen coding schemes could be applied to FIND films. Although home-based videos elicit more naturalistic play interactions than those filmed in lab, these films also tend to vary in quality; many home-based films have poor lighting, significant background noise, and may include other caregivers and/or children. Coding dyadic interactions between the target caregiver and child can be particularly challenging in these contexts, when distractors exist that were not present during the norming/standardization process for each measure. Our evaluation of feasibility included four primary questions:

1. Was it possible to use this coding scheme with brief (i.e., 5-10 minute) home-based free-play interactions?
2. In what situations would films be considered “uncodeable” using this coding instrument?
3. What percentage of our films (and percentage of time within each film) contain “uncodeable” interactions?
4. Can this scale be applied in multiple contexts? Would this coding instrument be valid in other languages and/or cultural contexts?

Ultimately, answers to these three primary considerations were synthesized into a single “yes” or “no” determination of feasibility. Coding instruments that were assigned a “no” rating for current feasibility underwent a secondary evaluation process to determine whether protocol adaptations (i.e. changes to the coding instrument or the film collection protocol) could improve feasibility. This secondary evaluation process is discussed in more depth in the Discussion.

**Coder reliability.**

Establishing inter-rater reliability (IRR) is essential to ensure consistency within coded data. For the purposes of calculating inter-rater reliability, 20% of the films in each coded dataset were double scored. Ultimately, our evaluation of inter-rater reliability led us to consider two primary questions in our overall evaluation of the measurement tool:

* 1. Can a team of coders reach an acceptable level of of inter-rater reliability during training for this instrument?
  2. Can the same team of coders achieve a similar level of inter-rater reliability in the context of a full dataset of FIND films?

As with our evaluation of feasibility, answers to these questions helped to create a final yes/no determination of whether it was possible to achieve reliability across home-based FIND films.

There is significant debate within the field about the best way to measure IRR (e.g. Hallgren, 2012; Jansen, Wiertz, Meyer, & Noldus, 2003; Lombard, Snyder-Duch, & Bracken, 2002; Stolarova, Wolf, Rinker, & Brielman, 2014). Many methodologists (e.g. Bakeman, 2000; Dewey, 1983, Hollenbeck, 1978; McHugh, 2012) advocate for Cohen’s Kappa (Cohen, 1960), an index of reliability that measures the percentage of concordance while accounting for the likelihood of chance agreement. This index is particularly valuable for rating scales with limited behavioral categories (e.g. ratings of 1-3), where the chance that coders will randomly select the same score is higher. Pearson’s correlation coefficient, commonly known as Pearson’s *r*, is also commonly used as an index of reliability, particularly for interval-based data that measures the duration and frequency of specific behaviors (Jansen et al., 2003). It is important to note that while researchers frequently draw conclusions about inter-rater agreement from correlation analyses (e.g. Hane et al., 2019; Roggman et al., 2013; Vilaseca et al., 2019), Pearson’s *r* does not specifically measure coder agreement (Kottner et al., 2011; Stolarova et al., 2014). Some methodologists who criticize the use of Cohen’s kappa and Pearson’s *r* advocate for the use of intra-class correlation coefficients (ICCs). ICCs offer significantly more flexibility in how IRR is calculated, offering the choice to evaluate variance across a large cohort of coders or to assess absolute differences between fixed coding pairs (Hallgren et al., 2012; Koo & Li, 2016). ICCs evaluate a composite of variance accounted by coder differences (inter-rater reliability) and variance within coded behaviors (intra-rater reliability); as such, ICCs are frequently used to evaluate inter-rater agreement for nominal and ordinal observations (Hallgren et al., 2012).

Across the literature, researchers have emphasized that methods for evaluating inter-rater reliability should vary based on the nature of the coding system. For example, calculations of IRR for a global rating scale should account for random chance (e.g. Hane et al., 2019; Roggman et al., 2013; Vilaseca et al., 2019), while a moment-to-moment system needs to consider the timing, duration, and sequence of coded behaviors (Jansen et al., 2003). Given the variability of coded behaviors in the instruments we selected (i.e. duration-based, frequency-based, and global codes), we conducted different IRR analyses for each coding scheme, matching the type of reliability analysis with the specific nature of the coding instrument. We also selected distinct reliability thresholds for each reliability index, using guidelines and cutoff values established across the IRR literature (e.g. Brennan & Silman 1992; Hallgren, 2012; Lombard et al., 2002; McHugh, 2012). Threshold values and indices used for each measure are reported in Table 1.

***WECS and SRS.***

Historically, methodologists have advocated for intra-class correlation (ICC) and Cohen’s kappa as the two most reliable methods for establishing inter-rater reliability within global coding instruments (Lombard et al., 2002). Accordingly, Hane and colleagues (2019) report ICCs in their validation research of the WECS video-scoring instrument, while other research groups report Cohen’s kappa to correct for chance agreement between coder scores (e.g. Pretzer et al., 2019). Based on this precedent, we report percent agreement, Cohen’s kappa, and intra-class coefficients in our evaluation of IRR for these two global coding schemes.

***Simple Interactions: Connection and Reciprocity*.**

For the SI subscales, which are scored at a moment-to-moment timescale, the large volume of coded variables (e.g. frequency, duration, time of onset and offset) suggests the need for other methodologies to evaluate consistency across coders. Since the crux of micro-social coding lies within the *timing and duration* of coded behaviors, we calculated reliability using a confusion matrix that takes both the duration and sequence of coded behaviors into account (see Jansen et al., 2003 for a full description of these calculations embedded into the Noldus Observer coding software). For these two SI subscales, we also calculated % match based only on the overall duration of each coded variable throughout the coded segment; since the SI scoring rubric ultimately weights each score based on overall duration within the film, this second reliability metric provides insight into whether these final weighted scores were reliable across coders.

***Conversational Turns*.**

Because the Conversational Turns scale tracked the frequency and duration of verbalizations throughout the film, we used a distinct reliability approach for this instrument. First, we calculated a percent match of utterance designations at 0.5 second intervals throughout each full coded segment. To maintain consistency with the way utterances were coded by raters, we calculated percent match values for each tier (parent vs. child utterances) separately. We then aggregated these scored utterances into 15 second bins and calculated Pearson’s *r* to determine the correlation between coders’ aggregated scores. To account for chance agreement between raters making binary (yes/no) decisions about utterances, we also calculated Cohen’s kappa values for both parent and child utterance tiers. These three reliability metrics (overall % match, aggregated correlations, and Cohen’s Kappa) have been used by other researchers coding time-sensitive language data (e.g. Bulgarelli & Bergelson, 2019; Mendoza & Fausey, 2019; Pretzer et al. 2019) and reflect a comprehensive approach to considering inter-rater coding reliability with data that is time-sensitive and behaviors that are relatively infrequent.

**Viability of inclusion in FIND implementation work.**

The third domain employed in the evaluation of coding measures assessed the relative strengths and weaknesses of each coding instrument in terms of implementation. Considering our primary aim to select coding measures that could be used for the ongoing evaluation of parenting within the context of community implementations of FIND, we targeted two primary questions:

1. Training - How challenging is it to train a coding team to reliability? How long does this training process take?
2. Time - How long does it take to code each video? What other time considerations are relevant to the timing of the coding process (e.g. data cleaning and analysis)?

Given the diversity of coding schemes selected for this study, it wasn’t possible to evaluate this domain using a dichotomous yes/no distinction. Rather, we present both qualitative and quantitative implementation considerations for each measure, including next steps for adapting, changing, and translating each instrument if necessary.

**Instrument Adaptation and Revision**

This evaluation process involved significant collaboration with the developers of coding instruments. Amie Hane (Williams College) and Martha Welch (Columbia University Medical Center) served as project consultants for our evaluation of the WECS, Judy Cameron (University of Pittsburgh) and Ian Murphy (University of Pittsburgh) consulted throughout the evaluation process for the Simple Interactions measure, and the FIND expert clinicians at the University of Oregon were instrumental in the training process for the SRS Scale. Conversational Turns was adapted from similar coding rubrics by our research team at the University of Oregon, designed specifically to complement the other coding schemes used in this study.

During the evaluation process, several coding instruments failed domain criteria (i.e. faced feasibility, reliability, or implementation challenges). Measures that faced one or more setbacks were reworked and revised in collaboration with the original development team, with the goal of creating an adapted version of the measure that could successfully be used to evaluate FIND films. This adaptation process involved fast-cycle iteration, weekly revision and evaluation meetings, and significant collaboration across institutions. We describe the revision process for specific measures in more depth in the Discussion section below.

For each instrument, we considered each evaluation criteria sequentially, determining feasibility first, followed by reliability, and finally implementation considerations. Coding instruments that failed one or more domain criteria were discussed in depth with the original developers to evaluate whether there was value in revising and adapting the measure to fit the specific context of FIND intervention films.

**Results**

Results for each coding instrument are described sequentially, addressing a) feasibility, b) reliability, and c) implementation viability for each measure. For coding instruments that failed any of the three domain criteria, the evaluation process was paused in order to assess whether the instrument should be adapted and revised. Therefore, we only report results from domains that were evaluated; coding measures that failed early in the evaluation process (e.g. feasibility) were not assessed for reliability and implementation considerations.

**WECS**

**Feasibility.**

In the context of our current study, we determined that the WECS was not feasible as a coding instrument to evaluate the FIND intervention. During initial conversations with the development team (see Hane et al., 2019), we discovered that the coding system required a minimum of 10 minutes of face-to-face interaction between the caregiver and child. This experimental design required the child to sit on the caregiver’s lap and/or to be strapped into a car seat or carrier that allowed uninterrupted face-to-face interaction with the caregiver. Further, this paradigm required two separate cameras to fully capture the facial expressions of the child and caregiver separately for coding. FIND films involve significantly more dynamic interaction during free-play, and are filmed from one angle only, making it difficult to see both faces at points throughout the film. Extensive discussions with the development team of the WECS yielded further concerns about the validity of adapting this coding scheme to fit a new context (free-play), compared to its standardized original form. We ultimately chose to terminate our evaluation of this coding scheme early in the process, focusing instead on evaluating, revising, and adapting the remaining three measures.

**SRS**

**Feasibility.**

Because the Serve and Return Scale was designed specifically to evaluate films from the FIND intervention, we had fewer concerns about the feasibility of this coding scheme in the context of home-based free-play interaction. This global coding scheme is significantly more flexible than the WECS, organizing specific items clear subscales based on the FIND 5 Elements. For items with lower base-rates in interaction (e.g. “Caregiver soothes or comforts the child when the child is distressed”), the coding scheme included a “No Opportunity” code, which could be used if that specific behavior didn’t occur within the film. This increased flexibility allowed for all FIND films to be coded, regardless of whether every type of interaction occurred within the 6-minute film. Further, because the coding scheme was rated globally (i.e. based on general impression of each item), every film that contained free-play interactions could be coded, regardless of film quality, lighting, or angle.

Because this coding scheme required a deep understanding of the linguistic content of each interaction, it was important that coders were fluent in the language spoken by caregivers and children in the films they were rating. Further, accurate coding of interactions across linguistic and cultural contexts requires careful and appropriate translation and adaptation (e.g. Beaton, Bombardier, Guillemin, & Ferraz, 2000; Sidani, Guruge, Miranda, Ford-Gilboe, & Varcoe, 2010; Sousa & Rojjanasrirat, 2011). Because the SRS instrument was only available in English, we only coded the English films from this dataset (n=51); further discussion about the translation and adaptation of coding measures is addressed in the Discussion.

**Reliability.**

Overall, the SRS scale did not pass our criteria for reliability. We evaluated three indices of inter-coder agreement – percent agreement, Cohen’s Kappa, and ICCs – in two contexts: a) during training and b) within the FIND ACF dataset. Although our video-coding teams were able to achieve high inter-rater reliability during training, this consistency was not maintained while coding films from the Denver ACF dataset.

During training, our coding team coded 69 videos from previous FIND coaching sessions, using both home-based and lab-based films of dyadic free-play interactions. By Week 9 of the training period, the SRS coding team was able to pass all three reliability thresholds, for percent agreement (>80%), kappa (>0.70), and ICC (>0.80). Reliability scores for training data were calculated by comparing individual coders’ scores on each film with gold standard “expert” ratings. Reliability statistics for the last week of training are reported as averages across all coded films assigned that week (n=8); these results are reported in Table 2.

Following training, coders scored the FIND films from Denver across the 3 trained subscales. While coding these FIND films, coders documented several significant concerns about contexts that had not appeared during training: several films had multiple children present, making it difficult to reliably code dyadic interaction since the presence of second or third child impacted the nature of the free-play. Further, film quality (e.g. background noise, poor lighting, obscured camera angles) impacted coders’ ability to hear conversations and accurately evaluate parent responses to their child’s serves. Finally, all of the film in the Denver dataset included 6 minutes of interaction, while most of the training videos were 10 minutes long. This change in film length contributed to challenges as coders adjusted their base rate frequencies to accurately score “low” (1), “medium” (2), and “high” (3) occurrences of specific behaviors (e.g. object naming, caregiver serves, and verbal praise). These differences between training film and the Denver FIND films translated to a heavy cost in inter-rater reliability: across double-coded films (n=35), coder percent agreement, ICCs, and Kappa values all fell below the desired reliability threshold. Results from these three reliability indices are reported in Table 3.

Notably, this drop in reliability impacted some films significantly more than others: while scores were consistently above threshold for all three reliability indices during training, inter-rater agreement varied widely across double-coded films in the Denver ACF dataset. This wide range in agreement indicates that some videos (e.g. videos with multiple children, films with poor audio or lighting quality) may have been harder to code accurately than others. These factors may have led to the widening range of inter-coder agreement scores rather than a straightforward drop in reliability across all films.

**Simple Interactions: Connection**

**Feasibility.**

The Connection subscale of the Simple Interactions instrument is designed for flexible use across dynamic parent-child interactions. Scored on a moment-to-moment (i.e. frame-by-frame) basis, this coding scheme utilizes a detailed glossary and complex flow chart that allows coders to quantify every possible situation within dyadic interaction. Given the dynamic nature of children’s play, this coding instrument also accounts for moments where interactions were “uncodeable”: frames where parent or child faces were obscured or where one dyadic partner was out of view. In these contexts, the quality of the video and angle of the camera prevent coders from being able to make accurate decisions using the flowchart. Coders are instructed to label these sections of film “No Code” in lieu of assigning a code based on inference. Because Connection was designed for use with hand-recorded films, the coding scheme operates on the underlying assumption that the videographer will adjust the filming angle to capture faces, maximizing the percentage of codeable film.

In this study, we found a wide range of film quality across the 176 FIND ACF films, yielding a broad range of “No Code” assignments (Range = 0 to 99.83%). 40% of films (n=71) contained at least one “No Code” assignment, and 15% of films (n=27) were given “No Code” labels for a majority of the film duration (No Code > 50% of coded film). In collaboration with the Simple Interactions development team at the University of Pittsburgh, we established a cutoff rule to define the minimum amount of codeable interaction to constitute a valid Connection score (min = 3 minutes of aggregated codable film). After defining this threshold, we calculated that 83.5% of our films (n=147) fit this criteria. Despite some lost data, these results indicate high codeability across the Denver ACF films and contributed to our overall determination that the Connection subscale was feasible for FIND films.

It should also be noted that the Connection instrument primarily requires coders to make decisions based on affect, synchrony, and physical interactions during play. As a result, the coding scheme does not require a deeper level understanding of the *content* of interactions, and can be coded regardless of language fluency. Because this coding instrument did not require translation or adaptation into Spanish, all 176 films from the Denver dataset were coded using this measure.

**Reliability.**

Calculation of inter-rater reliability across the Connection coding team was considered using two primary methods a) duration match and b) duration-sequence match. The first calculation (duration) considers how closely coders agreed on the overall percentage of time that four codes were assigned: CX, CY, CZ, and No Code, while the second reliability metric (duration-sequence) accounts for both the duration and the timing of each assigned code. Duration/sequence reliability is considered a stricter reliability threshold since it takes into account frequency, timing, and duration of each coded event.

To maintain consistency with the original version of Simple Interactions developed at the University of Pittsburgh, we established our training reliability threshold based on duration-based agreement between coders. We report 3 measures of inter-rater agreement: % agreement, Pearson’s r, and Kappa (See Table 1). Throughout the training period, connection coders were given weekly independent coding assignments and were asked to complete conferences in pairs to address discrepancies between their ratings. While coders were able to pass the reliability threshold for Pittsburgh “Gold Standard” films within the first 4 weeks of training, the context of FIND films provided an additional challenge. Caregivers and children in FIND films tended to be more active and moved freely throughout filmed interaction much more than caregivers in the Pittsburgh Gold Standard films. As such, determinations of “No Code” (e.g. when a child or caregiver’s face wasn’t fully visible) were more frequent, and our Connection coders needed several additional weeks to reach reliability thresholds for No Code assignments within films. Reliability indices averaged across 8 training videos coded during Week 8 are reported in Table 4, highlighting the point at which this coding team passed the reliability threshold for Connection.

Following training, coders scored the FIND films from the Denver ACF dataset using the same glossary, flowcharts, and independent coding procedures. As with SRS, coders documented several concerns about some contexts that had not appeared during training. Notably, films from the ACF dataset contained much higher percentages of “No Code” than training film, since poor lighting and filming angles made it difficult to see faces. These added dimensions of challenge impacted reliability, causing increased concern about the threshold of inference that could be used to determine facial expressions before assigning a score of “No Code.” Despite these challenges, our coding teams were able to achieve a high level of inter-rater reliability across both Duration and Duration/Sequence calculations, noted in Table 5. While kappa statistics fell below threshold for both duration and duration/sequence, our coding teams maintained high levels of matched agreement and high correlations (Pearson’s r) between coders across all double-coded films (n= 33).

**Implementation viability.**

The original training process for the Connection instrument took place over the course of 8 weeks, with each coder independently coding ~5 hours per week and attending coding meetings 1 hour per week. This equates to ~48 hours of training per coder to achieve full reliability within training videos. Of note, this training period coincided with the acquisition and setup of the Noldus Observer XT software system, which hosts all of the videos, coding schemes, and coding tools necessary for accurate Connection coding. Subsequent training cycles with coders learning the Connection instrument saw significantly reduced training needs; average training time for a second group of Connection coders was 5 weeks (~30 hours total), using Gold Standard films and Noldus templates that the first training group created.

After achieving reliability on training films, Connection coders reported an average of 60 minutes of coding per 6-minute free-play interaction; coding time varied significantly depending on film quality and participant facial expressions (range is estimated by coders to be between 30 and 90 minutes). Overall, these training and coding labor costs align with the well-established cost/benefits of microsocial coding schemes, representing a relatively large pull of resources to code a full dataset.

**Simple Interactions: Reciprocity**

**Feasibility.**

The Reciprocity subscale of the Simple Interactions instrument was originally designed to capture serve and return interactions in children ages 3-6 during dyadic interaction. It utilized the same glossary as the other two subscales (Connection and Progression), which limited the definitions of “serve” and “return” to verbal exchanges between caregivers and their children. Before starting the evaluation of the Reciprocity subscale, we fully adapted and revised the Reciprocity scale to redefine nonverbal “serves” and “returns” in the context of interaction with children 0-3. This new glossary (“Reciprocity 0-3”) can be found in Appendix D.

Because the adaptation process for this measure utilized FIND films to create the new glossary, the 0-3 version of the measure was designed with FIND intervention films in mind. Despite this advantage, our coding teams still ran into challenges within the Denver ACF films because of issues with film quality. For example, in order to capture gesture-based serve and return interactions, it is necessary to capture the faces and hands of both caregivers and children during play. 15-second segments where hands or faces are not clearly visible are assigned a label of “No Code,” since coders cannot accurately quantify serves and returns without this full visibility. Within the FIND ACF dataset, 98.0% (n=50) of the coded videos contained segments of “No Code” interaction because of film quality issues. By applying the feasibility threshold established with the research team at the University of Pittsburgh for Simple Interactions, we found that 90.2% (n=46) of these coded films were codeable (i.e. contained at least 3 minutes of codeable interaction). Given this high proportion of codeability, we determined that the Reciprocity measure was a feasible coding instrument for use with FIND intervention films.

Because the Reciprocity scale requires nuanced determinations that quantify serve and return interactions between parents and their children, this instrument requires a deep level of contextual understanding in order to accurately code interactions. As such, the Reciprocity scale in its current form can only be used to quantify films that contain interactions in English. Data reported here are based on the subset of English-only films from the Denver ACF dataset (n=51). Current efforts to adapt and translate the Reciprocity 0-3 instrument for use with Spanish-speaking families are described in more depth in the Discussion.

**Reliability.**

As with the Connection subscale of the Simple Interactions instrument, reliability for the Reciprocity scale was measured across two domains using the Noldus XT Observer software: a) Duration/Sequence and b) Duration. Within these two domains, we calculated three indices of IRR: % agreement, Pearson’s r, and Kappa.

Throughout the training phase of this measure, we utilized the same thresholds for reliability as we did for the Connection subscale. Notably, we used the same coding team that adapted the Reciprocity subscale from the original measure to code the Denver ACF dataset; therefore, it is difficult to parse the time these coders spent in training compared with time they were actively involved in measure development. Training reliability numbers are reported as averages of pairwise-comparisons applied across the full group of 5 coders. Table 6 shows reliability indices for films coded during the last week of training (n=4) before the team started coding the Denver ACF dataset.

As shown in Table 6, Reciprocity team reliability was hovering right at the threshold point for % agreement and Pearson’s *r* at the end of the training cycle. Across training films, kappa values did not meet reliability thresholds, indicating that some aspects of the matched agreement between coders reported by other indices may be overly influenced by random chance. These below-threshold reliability values were particularly concerning given that these evaluations were conducted after this coding team spent 4 months developing and training with the Reciprocity 0-3 glossary.

Because our independent coder reliability did not consistently meet threshold for reliability despite four months of intensive exposure to the coding rubric, we modified our coding protocol to include a convergent coding system for the Denver ACF films. With this new protocol, two coders would score each film independently and conduct a coding conference to address all discrepancies between their ratings. Using this system, two coders worked together to create a convergent code for every film. Although this protocol added significant coding time to the process (see Implementation), this method increased reliability across the coding team significantly. Reliability indices for both independent and convergent coded films are reported in Table 7, showing average pairwise comparisons across double-coded films. Independent reliability indices report averages of pairwise comparisons for all coders who coded each film (n=51); convergent reliability indices were calculated by having a separate pair of coders conference and double-score 20% of the convergent coded videos (n=9). Notably, averages for these double-conferenced films show extremely high reliability across conferencing pairs, indicating that the use of convergent codes was a successful (if time-consuming) way to reliably score films with the Reciprocity 0-3 instrument.

It is also important to note that while reliability averages for independently coded films hovered just below our reliability thresholds, there was significantly more variability in IRR across films scored by independent raters. Coders reported that conferences were particularly helpful during films of younger, pre-verbal infants, where labelling gesture-based serves could be especially challenging. Accurate coding using the Reciprocity subscale often required raters to infer the intentions behind infant movements. For example, a rater may need to decide if an infant turning her head is responding to an action from the caregiving (“return”) or expressing interest in a new toy (“serve”). Because of how challenging it can be to correctly classify these microsocial interactions, coders reported increased confidence in their scores when they had the opportunity to conference and work with others to create convergent scores.

**Implementation viability.**

Because this coding scheme was redesigned and modified from its original manualized version, it is difficult to report accurate “training time” data. The current coding team took ~4 months to redevelop the scale, working over 600 hours to hone and perfect the definitions of nonverbal serve and return and achieve reliability between coders. This new version (Reciprocity 0-3) has now been manualized, and a team at the University of Pittsburgh is currently in training. Based on previous experience with the Connection subscale of the Simple Interactions measure, we anticipate that it will take a similar amount of time for coders to achieve reliability: 4-6 weeks of independent coding practice with weekly feedback and conferencing.

The current version of the measure requires each video to be coded by two separate raters, who conference to finalize a convergent code. From start to finish, this process takes ~3 hours (1 hour per rater for coding, plus an additional hour for conferencing and convergent code finalization) per 6-minute film. Based on our reliability data, we can see that convergent coding allows video-coding teams to achieve high levels of IRR across all reliability indices; we will continue to explore ways to improve individual-coder reliability for future training protocols to allow for independent coders to finalize codes without the conference + convergent coding process, significantly decreasing the coding time required by this measure. In any case, it is important to note the significant time investment involved in micro-social coding schemes. Although moment-to-moment codes provide highly detailed data about the duration and frequency of specific behaviors and events, these microsocial coding instruments come at a significant labor cost.

**Conversational Turns**

**Feasibility.**

Because the CT coding scheme was designed by the author at the University of Oregon, we were able to build feasibility considerations into the coding design. As a result, the CT measure was designed for maximum feasibility and utility within the context of at-home free-play films. Although our coding teams did run into some challenges in coding (e.g. differentiating between speakers and/or musical toys, variable audio quality), 98.3% (n=173) of the Denver ACF films were codeable using this instrument. The 3 videos that were designated as “uncodeable” with the CT scale each had significant extenuating circumstances (e.g. audio playback difficulties, significant background noise) that would impact these films’ ability to be coded using any coding measure.

The CT measure was also specifically designed to quantify utterances using time-bound and *not* semantically-informed definitions. This allowed our coders to differentiate utterance boundaries with greater accuracy, since they weren’t relying on inference to determine when caregivers or children paused while they were talking. By defining utterance boundaries at time-based intervals (>0.25 seconds), we also alleviated the need for fluent bilingual coders, since coders were making judgements about the *length* of time and not the *meaning* of each utterance. Accordingly, all of the FIND ACF films were coded with CT (n=176), regardless of the family’s primary language.

**Reliability.**

Coding teams achieved high reliability using this instrument, measured three ways: a) percent match between coders at the 0.5 second level b) Cohen’s kappa and c) Pearson’s *r* calculated by aggregating utterance data into 15-second bins. Reliability was calculated separately for parent utterances and child utterances.

Because this coding instrument is significantly less complex than the other coding instruments evaluated in this study, coders were able to achieve high levels of reliability after just one training session. Reliability for this training film is reported in Table 8, averaging pair-wise comparisons of all coders who coded this film (n=6) when compared to an expert coder.

Consistent with the high IRR values found during training, coders maintained a high level of inter-rater agreement throughout the entirety of the FIND ACF dataset. Average IRR values for double-coded films (n=29) are reported in Table 9. Notably, while percent agreement values remained high for both parent and child utterances, we saw significant variability for correlation and kappa values within coded child utterances (SD = 0.15 for both indices). Specifically, coder agreement across these two indices dropped significantly for films with younger infants, where base rates for utterances were low. 6.9% (n=2) of the double coded films yielded Pearson’s *r* and kappa values <0.5 for child utterance ratings. For both of these films, coders rated very few (<20) child utterances across the full 6-minute interaction; when base rates of utterances are very low, slight differences in coder ratings can have a large impact on reliability ratings. Because our percent agreement rates a “match” if both coders agree on a binary rating (yes/no), coders can achieve high levels of agreement for a film that has few utterances (i.e. percent match includes matched silence). Both Kappa and Pearson’s *r* statistics only take into account positive matches for utterances – not matches for silence – systematically favoring coded utterances and ignoring larger sections of coded silence. Because this only impacted a small subset of our sample with extremely low child utterance counts, we considered our coding reliability to pass threshold for the Conversational Turns measure overall.

**Implementation viability.**

In comparison with the other coding instruments evaluated in this study, the Conversational Turns measure had the simplest implementation process. Coders who worked on this measure consistently met reliability thresholds within a one-day training session that included guided practice and one 10-minute film for independent practice. The ease of the training process for CT aligns with the measure’s inherent minimalism: coders follow three primary rules that define *what* utterances are, *when* they start and end, and *who* should be coded in each tier. The simplicity of this measure allows relatively naïve coders who have little experience with behavioral coding to reach high levels of mastery within a few hours.

Beyond training, the CT instrument requires a significant time commitment from coders. Given that coders are expected to code utterance boundaries within 0.25 second accuracy, this instrument requires significant attention to detail and often requires coders to re-watch the same segment of film several times with slower playback speeds. On average, coders reported coding times of 45-75 minutes per 6-minute film within the FIND ACF dataset. Coding times varied depending on the nature of the film (e.g. reported coding times are faster for films with younger infants who produce fewer utterances); on average, per-film coding durations aligned with the other micro-social instruments (Simple Interactions) that were evaluated in this study. In sum, while the CT measure involves a simple and relatively quick training process, this instrument is still highly labor-intensive because of the amount of time it takes expert raters to code individual films.

**Discussion**

This study offers a novel approach to the evaluation of video-coding instruments to measure caregiver-child interaction. We considered varied theoretical perspectives on responsive parenting – maternal sensitivity, emotional connection, contingent responsiveness – and selected four coding instruments that aligned with each of these different perspectives. As a part of the selection process, we also chose measures that represented varying levels of zoom; we evaluated two global coding schemes (WECS, SRS) and two microsocial instruments (SI, CT), comparing the viability and (dis)advantages of these varied methodologies. As a part of our evaluation process, we compared each measure across three criteria: a) feasibility b) reliability and c) viability for implementation.

**The Reliability/Flexibility Tradeoff**

Overall, our results highlight several important lessons about the process of coding dyadic interactions. First, the inherent specificity that allows video-coding instruments to be coded reliably (e.g. detailed glossaries, flow-charts, expert training films) limits the flexibility of the measure to adapt to new contexts. Our evaluation of the WECS provided insight into this challenge. In creating the WECS, Hane et al. (2019) chose to limit the scope of interactions to include only face-to-face interactions between a caregiver and child. Within this narrow setting, the authors developed, standardized, and validated the WECS as a viable measure of emotional connection between a caregiver and child. These limited parameters help the instrument maintain high levels of reliability – within and across coders – but also limit researchers’ ability to use the coding tool in other contexts (e.g. to evaluate free-play in FIND films).

Another example of this relationship between specificity, reliability, and flexibility across situations was visible in our evaluation of the Serve and Return Scale (SRS). In contrast with the narrow focus of the WECS, the SRS was intentionally developed with the flexibility to adapt to diverse free-play contexts within FIND films. However, by broadly assessing varied categories of interaction (e.g. support, encouragement, naming, caregiver initiated activities), the SRS failed to define specific behaviors with enough precision. When these broad definitions were applied to new contexts outside of what was seen during training, inter-rater reliability dropped below threshold. Broadly, these lessons learned about feasibility and reliability from the WECS and SRS measures help to characterize the delicate balance between instrument specificity, inter-rater reliability, and flexibility across contexts.

**Microsocial vs. Global Coding: Pros and Cons**

Our comparison of microsocial vs. global coding tools yielded evidence to support both sides of the methodology debate. Rather than identifying one method as superior to the other, this study highlights the relative strengths and weaknesses of each.

While global coding instruments offer significant advantages in efficiency – coding time typically aligns with the length of the coded film segment– these advantages come at the cost of increased training time and decreased inter-rater reliability. Because global coding instruments require raters to generalize their impressions across several minutes of interaction, scores can be heavily influenced by coders’ pre-existing schema for infant behaviors. For example, coders who have significant experience working with infants may form different impressions of dyadic interactions than coders who are relatively inexperienced with children this age. Since global ratings are so reliant on coders’ ability to make inferences about infant behaviors, the training period is a critical time for raters on a coding team to form consistent schema. Successful inter-rater reliability depends on a coding team’s ability to effectively identify consistent definitions of “low,” “medium,” and “high” across diverse interactions. As such, the training period for global rating scales is often extensive, exposing coders to a broad range of low/medium/high interaction types to solidify the consistency of these benchmark thresholds across the team. Our experience with the SRS highlights this trade-off well: coding teams took nearly 9 weeks to reach reliability on training film, and struggled to maintain reliability across FIND ACF films once coder thresholds for “low” “medium” and “high” were shifted in a new context.

In contrast, micro-social coding instruments zoom in on parent-child interactions, relieving pressure for coders to form subjective impressions of prolonged interactions, improving reliability, and increasing coding time. For coding schemes that rely on moment-to-moment decisions (e.g. Connection, CT), the training process involves a relatively straightforward clarification of coding rules and definitions. For example, coders learning the SI Connection measure needed to internalize a detailed glossary of interaction definitions and navigate a flow-chart to make coding decisions. Training involved developing fluency with Noldus Observer XT software and learning to identify a range of thresholds where coding behaviors changed, a process that took several weeks. Because training in microsocial instruments involves mastery of comprehensive coding rules (rather than the subjective alignment of global thresholds for “low” “medium” and “high” across raters), training time is entirely dependent on the complexity of the coding scheme. In contrast to the SI Connection teams, CT coders were able to reach mastery after a single training session because of the relative simplicity of the coding rules and the ELAN software system.

Beyond moment-to-moment rating scales, our evaluation of SI Reciprocity provided insight into the strengths and pitfalls of coding instruments that are scored across short-duration-intervals. In most ways, training for this type of coding resembles the training process for systems that code frame-by-frame. For example, the SI Reciprocity subscale used the same detailed glossary, and a complex decision flow chart similar to the one used for the SI Connection scale. However, beyond simply mastering accurate coding frame-by-frame, short interval coding requires coders to accurately track several moment-to-moment behaviors across a longer time interval. Within Reciprocity, this meant correctly identifying and keeping track of all serves and returns within 15-second intervals of dyadic interaction. This added layer of complexity introduced additional opportunities for coder errors, resulting in increased challenges to inter-rater reliability. For the Reciprocity scale in particular, we found that convergent coding (i.e. having two independent raters conference to create one final code) offered a promising solution to reliability challenges. Our coding team was able to achieve high levels of inter-agreement with convergent codes – at the cost of added training and overall coding time.

Ultimately, by coding the same dataset with four different coding instruments, this study allowed us to directly compare and quantify the relative strengths and weaknesses of different coding methodologies. The failure of some instruments across our evaluation criteria (WECS, SRS) and relative success of others (SI, CT) provides valuable insight for best practices in the evaluation of responsive caregiving. Overall, we found that microsocial coding instruments were most successful in reliably quantifying parent-child interactions across diverse contexts. While these coding systems (e.g. SI Connection, SI Reciprocity, CT) provide a promising opportunity to evaluate responsive caregiving behaviors within well resourced, large-scale studies, it is important to note the extensive time and labor costs involved in their implementation. In light of our failure to identify a flexible, cost-effective coding system that can be applied within community implementations of FIND, future efforts will build on key takeaways from this study in pursuit of this aim.

**Next Steps**

Despite the failure of our two global coding instruments to meet the criteria outlined in this study, we are still motivated to find a flexible and cost-effective (i.e. not time intensive) coding system that will allow us to evaluate community implementations of FIND at scale. Currently, our team is redeveloping the Serve and Return Scale, utilizing the expertise of the original SRS coding team to change the measure through a fast-cycle iterative process. In close collaboration with FIND expert clinicians who created the original measure in 2016, the SRS team has been involved in an ongoing development and evaluation cycle over the past 10 weeks to resolve some of the psychometric issues that led to poor IRR. We will continue this process over the coming months, incorporating aspects of micro-social frequency coding into the overall global rating scale, and directly applying key takeaways from this study into the SRS redevelopment process.

In evaluating the feasibility of each coding system across contexts, we also highlighted a need for quality coding systems that can be applied in other languages. While some moment-to-moment coding measures define interaction boundaries without requiring linguistic comprehension (e.g. CT, SI Connection), most coding schemes based on short-interval codes or global impressions require semantic understanding of interaction content. In this study, the SRS and SI Reciprocity scales could not be applied to our Spanish films without additional translation and adaptation efforts. Over the next several months, our expert Reciprocity team will conduct a full translation and adaptation of the measure, following guidelines outlined by Beaton et al. (2000). This process – which involves translation, back-translation, and careful adaptation of coding rules to account for cultural differences in caregiver-infant interactions – will be completed by a team of fully bilingual and bicultural expert coders from diverse Spanish-speaking backgrounds (n=6). The current SRS expert coding team (n=6) is also comprised of bilingual/bicultural expert coders who plan to complete this process after the SRS 2.0 measure has been finalized and evaluated.

Finally, an important next step in the assessment process for all of these coding instruments involves evaluating sensitivity to change. Given our primary aim to evaluate how the FIND intervention *changes* parenting behaviors, it is critical that the coding instruments we select demonstrate this capacity for change. The FIND ACF films that were coded for this study were drawn from a large randomized controlled trial of FIND. Next steps for assessing sensitivity to change will involve unblinding coded films and running regression analyses to assess for intervention effects. This evaluation will also systematically assess for floor and ceiling effects that may be present during “pre-intervention” interactions (for methods, see Lotzin et al., 2015).

Ultimately, this study offers a direct comparison of different video-coding methodologies by evaluating four different coding instruments across the same dataset of free-play films. This evaluation highlights several of the challenges embedded in behavioral coding methods: applying coding schemes across diverse interaction contexts, achieving high levels of inter-rater reliability, and balancing the quantity and quality of coded data with labor costs (i.e. training and coding time). Beyond the dimensions directly measured in this study, there are several other considerations relevant to selecting an ideal coding measure for capturing behavioral intervention effects. Future directions for this research will consider the validity of these coding instruments in addition to their direct alignment with the FIND theory of change.

References

Ainsworth, M. D. S. (1969). Maternal sensitivity scales. *Power, 6*, 1379–1388.

Ainsworth, M. D. S., Blehar, M. C., Waters, E., & Wall, S. (1978). *Patterns of attachment: A psychological study of the strange situation.* Hillsdale, NH: Erlbaum.

Adamson, L. B., Kaiser, A. P., Tamis-Lemonda, C. S., Owen, M. T., & Dimitrova N. (2019). The developmental landscape of early parent-focused language intervention. *Early Childhood Research Quarterly*, 50(1), 59-67.

Akiva, T., Li, J., Martin, K., Horner, C., & McNamara, A. (2017). Simple Interactions: Piloting a strengths-based and interaction-based professional development intervention for out-of-school time programs. *Child & Youth Care Forum, 46*(3), 285-305.

Alexander, J. F., Newell, R. M., Robbins, M. S., & Turner, C. W. (1995). Observational coding in family therapy process research. *Journal of Family Psychology, 9,* 355-365.

Aspland, H. & Gardner, F. (2003). Observational measures of parent-child interaction: An introductory review. *Child and Adolescent Mental Health, 8*, 136-143.

Bakeman, R. (2000). Behavioral observation and coding. In H. T. Reis & C. M. Judge (Eds.), *Handbook of research methods in social and personality psychology* (pp. 138-159). New York: Cambridge University Press.

Bakermans-Kranenburg, M.J. Steele, H., Zeanah, C. H., Muhamedrahimov, R. J., Vorria, P., Dobrova-Krol, N. A., … Gunnar, M. R. (2011). Attachment and emotional development in institutional care: Characteristics and catch-up. *Monographs of the Society for Research in Child Development, 76*(4), 62-91.

Bandura, A. (2001). Social cognitive theory: An agentic perspective. *Annual Review of Psychology, 52,* 1-26.

Bardack, S., Herbergs, J. E., Obradovic, J. (2017). Unique contributions of dynamic versus global measures of parent-child interaction quality in predicting school adjustment. *Journal of Family Psychology, 31*(6), 649-658.

Beaton, D. E., Bombardier, C., Guillemin, F., & Ferraz, M. B. (2000). Guidelines for the process of cross-cultural adaptation of self-report measures. *SPINE, 25*(24), 3186-3191.

Bernard, K., Dozier, M., Bick, J., & Gordon, M. K. (2015). Intervening to enhance cortisol regulation among children at risk for neglect: results of a randomized clinical trial. *Developmental Psychopathology, 27*, 829-841.

Bernard, K., Hostinar, C. E., & Dozier, M. (2015b). Intervention effects on diurnal cortisol rhythms of Child Protected Services-referred infants in early childhood: Preschool follow-up results of a randomized clinical trial. *The Journal of the American Medical Association Pediatrics, 169*(2), 112-119.

Bernier, A., Carlson, S. M., & Whipple, N. (2010). External regulation to self-regulation: early parenting precursors of young children’s executive functioning. *Child Development, 81*(1), 326-339.

Bick, J., & Nelson, C. A. (2016). Early adverse experiences and the developing brain. *Neuropsychopharmacology, 41*(1), 177-196.

Bigelow, A. E., MacLean, K., Proctor, J., Myatt, T., Gillis, R., & Power, M. (2010). Maternal sensitivity throughout infancy: Continuity and relation to attachment security. *Infant Behavior and Development, 33*(1), 50–60.

Biringen, Z., Robinson, J. L., & Emde, R. N. (1998). *Emotional Availability Scale*s (3rd ed.). Fort Collins, CO: Department on Human Development and Family Studies, Colorado State University. Unpublished manual.

Blair, C., Granger, D. A., Willoughby, M., Mills-Koonce, R., & Cox, M. (2011). Salivary cortisol mediates the effects of poverty and parenting on executive functions in early childhood: cortisol and cognition. *Child Development, 82*(6), 1970-1984.

Blaisdell, K. N., Imhof, A. M., Fisher, P. A. (2019). Early adversity, child neglect, and stress neurobiology: From observations of impact to empirical evaluations of mechanisms. *International Journal of Developmental Neuroscience, 78,* 139-146.

Bohr, Y., Putnick, D. L., Lee Y. & Bornstein, M. H. (2018). Evaluating caregiver sensitivity to infants: Measures matter. *Infancy, 23*(5), 730-747.

Bornstein, M. H., Putnick, D. L., Cote, L. R., Haynes, M., & Suwalsky, J. T. D. (2015). Mother-infant contingent vocalizations in 11 countries. *Psychological Science, 26*(8), 1272-1284.

Bornstein, M. H., & Tamis-LeMonda, C. S. (1989). Maternal responsiveness and cognitive development in children. In M. H. Bornstein (Ed.), *New directions for child development, No. 43: The Jossey-Bass social and behavioral sciences series. Maternal responsiveness: Characteristics and consequences* (p. 49–61). San Francisco: Jossey-Bass.

Bornstein, M. H. & Tamis-LeMonda, C. S. (1997). Mothers’ responsiveness in infancy and their toddlers’ attention span, symbolic play, and language comprehension: Specific predictive relations. *Infant Behavior and Development, 20,* 283-296.

Bornstein, M. H., Tamis-LeMonda, C. S., Hahn, C. S., & Haynes, O. M. (2008). Maternal responsiveness to young children at three ages: Longitudinal analysis of a multidimensional, modular, and specific parenting construct. *Developmental Psychology, 44*(3), 867-74.

Bos, K. J., Fox, N., Zeanah, C. H., & Nelson, C. (2009). Effects of early psychosocial deprivation on the development of memory and executive function. *Frontiers in Behavioral Neuroscience, 3*, 16.

Brady-Smith, C., O’Brien, C., Berlin, L., Ware, A., & Brooks-Gunn, J. (1999). *Child-Parent Interaction Rating Scales for the Three-Bag Assessment 24-Month Wave*. National Center for Children and Families, Washington D. C. Unpublished manual.

Brennan, P., & Silman, A. (1992). Statistical methods for assessing observer variability in clinical measures. *British Medical Journal, 304,* 1491- 1494.

Bretherton, I. (2013). Revisiting Mary Ainsworth’s conceptualization and assessments of maternal sensitivity- insensitivity. *Attachment & Human Development, 15*, 460–484.

Bruce, J., Fisher, P. A., Pears, K. C., & Levine, S. (2009). Morning cortisol levels in preschool-aged foster children: Differential effects of maltreatment type. *Developmental Psychobiology, 51*(1), 14-23.

Bruce, J., Gunnar, M. R., Pears, K. C., & Fisher, P. A. (2013). Early adverse care, stress neurobiology, and prevention science: Lessons learned. *Prevention Science, 14,* 247-256.

Bulgarelli, F. & Bergelson, E. (2019). Look who’s talking: A comparison of automated and human-generated speaker tags in naturalistic day-long recordings. *Behavior Research Methods.* Advance online publication.

Carlson, V., Cicchetti, D., Barnett, D., & Braunwald, K. (1989). Disorganized/disoriented attachment relationships in maltreated infants. *Developmental Psychology, 25*(4), 525-531.

Caspi, A., Moffitt, T. E., Morgan, J., Rutter, M., Taylor, A., Arsenault, L., … Polo-Tomas, M. (2004). Maternal expressed emotion predicts children’s antisocial behavior problems. Using monozygotic twin differences to identify environmental effects on behavioral development. *Developmental Psychology, 40,* 149-161.

Cassidy, J. & Shaver, P. R. (Eds.) (2018). Handbook of attachment: Theory, research, and clinical applications (3rd ed.). New York, NY: The Guilford Press.

Clarke, A. S., Kraemer, G. W., Kupfer, D. J. (1998). Effects of rearing condition on HPA axis response to fluoxetine and desipramine treatment over repeated social separations in young rhesus monkeys. *Psychiatry Research, 79*(2)*,* 91-104.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational & Psychological Measurement, 20,* 37-46.

Davidov, M., & Grusec, J. E. (2006). Untangling the links of parental responsiveness to distress and warmth to child outcomes. *Child Development, 77*(1), 44-58.

De Bellis, M. D. (2005). The psychobiology of neglect. *Child Maltreatment, 10*(2), 150-172.

De Wolff, M. S., & van IJzendoorn, M. H. (1997). Sensitivity and attachment: A meta-analysis on parental antecedents of infant attachment. *Child Development, 68*, 571–591.

Demaio, L. J. (1982). Conversational turn-taking: A salient dimension of children’s language learning. *Speech and Language, 8,* 159-190.

Deans, C. L. (2018). Maternal sensitivity, its relationship with child outcomes, and interventions that address it: a systematic literature review. *Early Child Development and Care.* Advance online publication.

Dewey, M. E. (1983). Coefficients of agreement. *British Journal of Psychiatry, 143,* 487-489.

Dishion, T. J., Gardner, K., Patterson, G. R., Reid, J. B., & Thibodeaux, S. (1983). The family process code: A multidimensional system for observing family interaction. *Unpublished coding manual. Available from Oregon Social Learning Center,* 1-207.

Dishion, T. J., Mun, C. J., Tein, J. Y., Kim, H., & Shaw, D. S. (2017). The validation of macro and micro observations of parent-child dynamics using the relationship affect coding system in early childhood. *Prevention Science, 18*(3), 268-280.

Dozier, M., Manni, M., Gordon, M. K., Peloso, E., Gunnar, M. R., Stovall-McClough, K., … & Levine, S. (2006). Foster children’s diurnal production of cortisol: An exploratory study. *Child Maltreatment, 11*(2), 189-197.

Dozier, M., Peloso, E., Lewis, E., Laurenceau, J., & Levine, S. (2008). Effects of an attachment-based intervention on the cortisol production of infants and toddlers in foster care. *Developmental Psychopathology, 20*(3), 845-859.

Dozier, M., Peloso, E., Lindheim, O., Gordon, M. K., Manni, M., Sepulveda, S., & Ackerman, J. (2006). Developing evidence-based interventions for foster children: an example of a randomized clinical trial with infants and toddlers. *Journal of Social Issues, 62*(4), 767-785.

Dunst, C. J. & Kassow, D. Z. (2008). Caregiver sensitivity, contingent social responsiveness, and secure attachment. *Journal of Early and Intensive Behavior Intervention, 5*(1), 40-56.

Dunst, C. J., Lowe, L. W., Bartholomew, P. C. (1990). Contingent social responsiveness, family ecology, and infant communicative competence. *National Student Speech Language Hearing Association Journal, 17*, 39-49.

Eddy, J. M., Dishion, T. J., & Stoolmiller, M. (1998). The analysis of intervention change in children and families: Methodological and conceptual issues embedded in intervention studies. *Journal of Abnormal Child Psychology, 26,* 53-69.

Edmiston, E., Wang, F., Mazure, C., Guiney, J., Sinha, R., Mayes, L., & Blumberg, H. (2011). Corticostriatal-limbic gray matter morphology in adolescents with self-reported exposure to childhood maltreatment. *Archives of Pediatric and Adolescent Medicine, 165*(12), 1069-1077.

Egeland, B., & Sroufe, L. (1981). Attachment and early maltreatment. *Child Development, 52*(1), 44-52.

Egeland, B., Sroufe, A., & Erickson, M. (1983). The developmental consequence of different patterns of maltreatment. *Child Abuse and Neglect, 7*(4)*, 459-469.*

Eisenberg, N., Zhou, Q., Spinrad, T. L., Valiente, C., Fabes, R. A., & Liew, J. (2005) Relations among positive parenting, children’s effortful control, and externalizing problems. A three-wave longitudinal study. *Child Development, 76,* 1055-1071.

Eluvathingal, T. J., Chugani, H. T., Behen, M. E., Juhasz, C., Muzik, O., Maqbool, M., … & Makki, M. (2006). Abnormal brain connectivity in children after early severe socioemotional deprivation: A diffusion tensor imaging study. *Pediatrics, 117*(6), 2093-2100.

Eyberg, S. M., Duke, M., McDiarmid, M., Boggs, S., Robinson, E., & Washington, E. (2004). *Dyadic parent-child interaction coding system third edition.* Unpublished manuscript. Gainesville: University of Florida.

Eyberg, S. M., Nelson, M. M., & Boggs, S. R. (2008). Evidence-based psychosocial treatments for children and adolescents with disruptive behavior. *Journal of Clinical Child and Adolescent Psychology, 37*(1), 215-237.

Fagan, M. A., Frosch, C. A., Middlemiss, W., LaCoursiere, J. N., Owen, M. T., Hane, A. A., & Welch, M. G. (2019). The practical utility of the Welch Emotional Connection Screen for rating parent-infant relational health. *Infancy, 24,* 881-892.

Farah, M. J., Betancourt, L., Shera, D. M., Savage, J. H., Gianetta, J. M., Brodsky, N. L., Malmud, E. K., & Hurt, H. (2008). Environmental stimulation, parental nurturance and cognitive development in humans. *Developmental Science, 11*(5), 793-801.

Farrell, A. K., Simpson, J. A., Carlson, E. A., Englund, M. M., & Sung, S. (2017). The impact of stress at different stages of life on physical health and the buffering effects of maternal sensitivity. *Health Psychology, 36*(1), 35-44.

Fisher, P. A., Beauchamp, K. G., Roos, L. E., Noll, L. K., Flannery, J., & Delker, B. C. (2016). The Neurobiology of Intervention and Prevention. *Annual Review of Clinical Psychology, 12*, 331-357.

Fisher, P. A., Berry, M., Howell, K. R., & Noll, L. (2016). *Serve and Return Observational Rating Scale.* Stress Neurobiology and Prevention Research Laboratory, University of Oregon, Eugene, OR. Unpublished measure available upon request from the authors.

Fisher, P. A., Frenkel, T. I., Noll, L. K., Berry, M., & Yockelson, M. (2016). Promoting healthy child development via a two-generation translational neuroscience framework: The Filming Interactions to Nurture Development video coaching program. *Child Development Perspectives, 10*(4), 251-256.

Fisher, P. A., & Gilliam, K. S. (2012). Research into theory into practice: An overview of family based interventions for child antisocial behavior developed at the Oregon Social Learning Center. *Clinica y Salud, 23*(3), 247-259.

Fisher, P. A., & Kim, H. K. (2007). Intervention effects on foster preschoolers’ attachment-related behaviors from a randomized trial. *Prevention Science: The Official Journal of the Society for Prevention Research, 8*(2), 161-170.

Fisher, P. A., Stoolmiller, M., Gunnar, M. R., & Burraston, B. O. (2007). Effects of a therapeutic intervention for foster preschoolers on diurnal cortisol activity. *Psychoneuroendocrinology, 32*, 892-905.

Flannery, J. E., Beauchamp, K. G. & Fisher, P. A. (2017). The role of social buffering on chronic disruptions in quality of care: evidence from caregiver-based interventions in foster children. *Social Neuroscience, 21*(1), 86-91.

Floyd, F. J. (1989). Segmenting interactions: Coding units for assessing marital and family behaviors. *Behavioral Assessment, 11,* 13-29.

Fox, N. A., Almas, A. N., Degnan, K. A., Nelson, C. A., & Zeanah, C. H. (2011). The effects of severe psychosocial deprivation and foster care intervention on cognitive development at 8 years of age: Findings from the Bucharest Early Intervention Project. *Journal of Child Psychology and Psychiatry, 52*(9), 919-928.

Frick, M. A., Forslund, T., Fransson, M., Johansson, M., Bohlin, G., & Brocki, K. C. (2017). The role of sustained attention, maternal sensitivity, and infant temperament in the development of early self-regulation. *British Journal of Psychology, 109*(2), 227-298.

Fries, A. B., Shirtcliff, E. A., & Pollak, S. D. (2008). Neuroendocrine dysregulation following early social deprivation in children. *Developmental Psychobiology, 50*(6), 588-599.

Frosch, C. A., Fagan, M. A., Middlemiss, W., Chang, M. Hane, A. A., & Welch, M. G. (2019). Validation study showed that ratings on the Welch Emotional Connection Screen at infant six months are associated with child behavioural problems at age three years. *Acta Paediatricia, 108*(5), 889-895.

Gardner, F. (2000). Methodological issues in the direct observation of parent-child interaction: Do observational findings reflect the natural behavior of participants? *Clinical Child and Family Psychology Review, 3,* 185-198.

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Oller, D. K., … Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology, 26*(2), 248-265.

Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). (Baby)Talk to me: The social context of infant-directed speech and its effects on early language acquisition. *Current Directions in Psychological Science, 24*(5), 339-344.

Guiliani, N. R., Beauchamp, K. G., Noll, L. K., & Fisher, P. A. (2019). A preliminary study investigating maternal neurocognitive mechanisms underlying a child-supportive parenting intervention. *Frontiers in Behavioral Neuroscience, 13*(16), 1-14.

Gunnar, M R. (1998). Quality of early care and buffering of neuroendocrine stress reactions: Potential effects on the developing human brain. *Preventative Medicine, 27*(2), 208-211.

Gunnar, M. R., Morison, S. J., Chisholm, K., & Schuder, M. (2001). Salivary cortisol levels in children adopted from Romanian orphanages. *Developmental Psychopathology, 13,* 611-628.

Hallgren, K. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology, 8*(1), 23-24.

Hane, A., LaCoursiere, J. N., Mitsuyama, M., Wieman, S., Ludwig, R., Kwon, K. Y., … Welch, M. G. (2019). The Welch Emotional Connection Screen: validation of a brief mother-infant relational health screen. *Acta Paediatrica, 108*(4), 615-625.

Hanson, J. L., Adluru, N., Chung, M. K., Alexander, A. L., Davidson, R. J., & Pollak, S. D. (2013). Early neglect is associated with alterations in white matter integrity and cognitive functioning. *Child Development, 84*(5), 1566-1578.

Harlow, H. R., Dodsworth, R. O., & Harlow, M. K. (1965). Total social isolation in monkeys. *Preceedings of the National Academy of Sciences, 54*(1), 90-97.

Harlow, H. F., & Suomi, S. J. (1971). Social recovery by isolation-reared monkeys. *Proceedings of the National Academy of Sciences, 68*(7), 1534-1538.

Hawes, D. J., Dadds, M. R., & Pasalich, D. S. (2013). Observational coding strategies. In J. S. Comer & P. C. Kendall (Ed.), *The Oxford Handbook of Research Strategies for Clinical Psychology* (pp. 120 – 142). New York: Oxford University Press.

Hesse, E., & Main, M. (2000). Disorganized infant, child, and adult attachment: Collapse in behavioral and attentional strategies. *Journal of the American Psychoanalytic Association, 48*(4), 1097-1127.

Hirsh-Pasek, K. & Burchinal, M. (2006). Mother and caregiver sensitivity over time: Predicting language and academic outcomes with variable- and person-centered approaches. *Merrill-Palmer Quarterly, 52*(3), 449-485.

Hirschmann, N., Deimann, P., Schmelzer, M., & Pietschnig, J. (2018). Reliable and valid coding of thin slices of video footage: Applicability to the assessment of mother-child interactions. *Journal of Psychopathology and Behavioral Assessment, 40*(2), 249-258.

Hollenbeck, A. R. (1978). Problems of reliability in observational data. In G. P. Sackett (Ed.) *Observing behavior: Vol 12. Data collection and analysis methods* (pp.79-98). Baltimore: University Park Press.

Holler, J., Kendrick, K. H., Casillas, M., & Levinson, S. C. (2016). Editorial: Turn-taking in human communicative interaction. *Frontiers in Psychology,* 6-1919.

Hubbs-Tait, L., Culp, A. M., Culp, R. E., & Miller, C. E. (2002). Relation of maternal cognitive stimulation, emotional support, and intrusive behavior during Head Start to children’s kindergarten cognitive abilities. *Child Development, 73*(1), 110-131.

Ivy, A. S., Brunson, K. L., Sandman, C., Baram, T. Z. (2008). Dysfunctional nurturing behavior in rat dams with limited access to nesting material: A clinically relevant model for early-life stress. *Neuroscience, 154*(3), 1132-1142.

Jansen, R. G., Wiertz, L. F., Meyer, E. S., Noldus, L. P. J. (2003). Reliability analysis of observational data: problems, solutions, and software implementation. *Behavior Research Methods, Instruments, and Computers, 35*(3), 391-399.

Juffer, F., Bakermans-Kranenburg, M., van IJzendoorn, M. H. (Eds.) (2008). *Promoting positive parenting: An attachment-based intervention.* New York, NY: Taylor & Francis.

Keller, H., Lohaus, A., Volker, S., Cappenberg, M., & Chasiotis, A. (1999). Temporal contingency as an independent component of parenting behavior. *Child Development, 70,* 474-485.

Keller, H., Lohaus, A., Volker, S., Elben, C, & Ball, J. (2003). Warmth and contingency and their relationship to maternal attitudes toward parenting. *The Journal of Genetic Psychology, 164*(3), 2003.

Kemp, S. P., Marcenko, M. O., Lyons, S. J., & Kruzich, J. M. (2014). Strength-based practice and parental engagement in child welfare services: An empirical examination. *Children and Youth Services Review, 47,* 27-35.

Kim-Cohen, J., Moffit, T. E., Caspi, A., & Taylor, A. (2004). Genetic and environmental processes in young children’s resilience and vulnerability to socioeconomic deprivation. *Child Development, 75*(3), 651-668.

Kochanska G., Forman D. R., & Coy K. C. (1999). Implications of the mother-child relationship in infancy for socialization in the second year of life. *Infant Behavior & Development, 22*, 249–265.

Koo, T. K. & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163.

Koss, K. J., Hostinar, C. E., Donzella, B., & Gunnar, M. R. (2014). Social deprivation and the HPA axis in early development. *Psychoneuroendocrinology, 50,* 1-13.

Kottner, J., Audige, L., Brorson, S., Donner, A., Gajewski, B. J., Hrobjartsson A. … & Streiner, D. L. (2011). Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Journal of Clinical Epidemiology, 64*(1), 96-106.

Kreppner, J. M., O’Connor, T. G., Rutter, M., & English and Romanian Adoptees Study Team. (2001). Can inattention/overactivity be an institutional deprivation syndrome? *Journal of Abnormal Child Psychology, 29*(6), 513-528.

Kuchirko, Y. (2017). On differences and deficits: A critique of the theoretical and methodological underpinnings of the word gap. *Journal of Early Childhood Literacy, 19*(4), 533-562.

Landry, S. H., Smith, K. E., & Swank, P. R. (2003). The importance of parenting during early childhood for school-age development. *Developmental Neuropsychology, 24(*2-3), 559-591.

Landry, S. H., Smith, K. E., & Swank, P. R. (2006). Responsive parenting: Establishing early foundations for social, communication, and independent problem-solving skills. *Developmental Psychology, 42*(4), 627-642.

Landry, S. H., Smith, K. E., Swank, P. R., Assel, M. A., & Vellet, S. (2001). Does early responsive parenting have a special importance for children’s development or is consistency across early childhood necessary? *Developmental Psychology, 37*(3), 387-403.

Leerkes, E. M., Blankson, N., & O’Brien, M. (2009). Differential effects of maternal sensitivity to infant distress and non-distress on social-emotional functioning. *Child Development, 80*(3), 762-775.

Li, J. (2014). Simple interactions tool. The Fred Rogers Center. Retrieved from <http://www.simpleinteractions.org/the-si-tool.html>.

Li, J. & Julian, M. (2012). Developmental relationships as the active ingredient: A unifying working hypothesis of “what works” across intervention settings. *American Journal of Orthopsychiatry, 82*(2), 157-166.

Lohaus, A., Keller, H.. Ball, J., Elben, C., & Volker, S. (2001). The concept of maternal sensitivity: Components and relations to warmth and contingency. *Parenting: Science and Practice, 1*, 267-284.

Loman, M. M., Wiik, K. L., Frenn, K. A., Pollak, S. D., & Gunnar, M. R. (2009). Post-institutionalized children’s development: Growth, cognitive, and language outcomes. *Journal of Developmental and Behavioral Pediatrics, 30*(5), 426-434.

Lombard, M., Snyder-Duch, J., Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research, 28*(4), 587-604.

Lotzin, A., Lu, X., Kriston, L., Schiborr, J., Musal, T., Romer, G., & Ramsauer, B. (2015). Observational tools for measuring parent-infant interaction: A systematic review. *Clinical Child and Family Psychology Review, 18*(2), 99-132.

Ludy-Dobson, C. R. & Perry, B. P. (2010). The role of healthy relational interactions in buffering the impact of childhood trauma. In E. Gil (Ed.), *Working with children to heal interpersonal trauma: The power of play.* New York, NY: The Guilford Press*.*

Maestripieri, D., McCormack, K., Lindell, S. G., Higley, J. D., Sanchez, M. M. (2006). Influence of parenting style of the offspring’s behavior and CSF monoamite metabolite levels in crossfostered and noncrossfostered female rhesus macaques. *Behavioral Brain Research, 175,* 90-95.

Maheu, F. S., Dozier, M., Guyer, A. E., Mandell, D., Peloso, E., Poeth K., … & Ernst, M. (2010). A preliminary study of medial temporal lobe function in youths with a history of caregiver deprivation and emotional neglect. *Cognitive Affective and Behavioral Neuroscience, 10*(1), 34-49.

Margolin, G., Oliver, P. H., Gordis, E. B., O’Hearn, H. G., Medina, A. M., Ghosh, C. M., & Morland, L. (1998). The nuts and bolts of behavioral observation of marital and family interaction. *Clinical Child and Family Psychology Review, 1,* 195-213.

Markman, H. J., Leber, B. D., Cordova, A. D., & St. Peters, M. (1995). Behavioral observation and family psychology: Strange bedfellows or happy marriage? Comment on Alexander et al. (1995). *Journal of Family Psychology, 9,* 371-379.

Markman, H. J. & Notarius, C. I. (1987). Coding marital and family interaction: Current status. In T. Jacob (Ed.) *Family interaction and psychopathology: Theories, methods, and findings* (pp. 329-375). New York: Plenum Press.

Marshall, P. J., Fox. N. A., & the BEIP Core Group. (2004). A comparison of the electroencephalogram between institutionalized and community children in Romania. *Journal of Cognitive Neuroscience, 16*(8), 1327-1338.

Mathis, E. T., & Bierman, K. L. (2015). Dimensions of parenting associated with child prekindergarten emotion regulation and attentional control in low-income families. *Social Development, 24*(3), 601-620.

McElwain, N. L., & Booth-LaForce, C. (2006). Maternal sensitivity to infant distress and non-distress as predictors of infant-mother attachment security. *Journal of Family Psychology, 20*, 247–255.

McHugh, M. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276-282.

Mehta, M. A., Golembo, N. I., Nosarti, C., Colvert, E., Mota, A., Williams, S. C. R., … & Sonuga-Barke, E. J. S. (2009). Amygdala, hippocampal and corpus callosum size following severe early institutional deprivation: the English and Romanian adoptees study pilot. *Journal of Child Psychology and Psychiatry, 50*(8), 943-951.

Meins, E., Fernyhough, C., Fradley, E., & Tuckey, M. (2001). Rethinking maternal sensitivity: Mothers’ comments on infants’ mental processes predict security of attachment at 12 months. Journal of Child Psychology and Psychiatry, 42, 637–648.

Mendoza, J. K., & Fausey, C. M. (2019, November 23). Manuscript: Everyday music in infancy. Retrieved from osf.io/j93n2

Morris, A. S., Criss, M. M., Silk, J. S., & Houltberg, B. J. (2017). The impact of parenting on emotion regulation during childhood and adolescence. *Child Development Perspectives, 11*(4), 233-238.

Murphy, I., Laws, M., Miragaia, A. S., DeMand, A., Tummala, S., Wanner, J., … Cameron, J. L. (2018). *Simple Interactions Video-Coding Tool.* University of Pittsburgh, Pittsburgh, PA. Unpublished measure available upon request from the authors.

Nelson, E. E., & Winslow, J. T. (2009). Non-human primates: Model animals for developmental psychopathology. *Psychopharmacology, 34,* 90-105.

Nelson, C. A., Zeanah, C. H., Fox, N. A., Marshall, P. J., Smyke, A. T., & Guthrie, D. (2007). Cognitive recovery in socially deprived young children: The Bucharest Early Intervention Project. *Science, 318,* 1937-1940.

Parker, S. W., Nelson, C. A., & the BEIP Core Group. (2005a). An event-related potential study of the impact of institutional rearing on face recognitions. *Development and Psychopathology, 17*(3), 621-639.

Parker, S. W., Nelson, C. A., & the BEIP Core Group (2005b). The impact of early institutional rearing on the ability to discriminate facial expressions of emotion: An event-related potential study. *Child Development, 76*(1), 54-72.

Patterson, G. R. (1982). A social learning approach: Vol. 3: Coercive family process. *Behavior Therapy, 15*(1), 121-127.

Patterson, G. R., & Reid, J. B. (1984). Social interactional processes within the family: The study of moment-by-moment family transactions in which human social development is embedded. *Journal of Applied Developmental Psychology, 5,* 237-262.

Patterson, G. R., Reid, J. B., & Dishion, T. J. (1992). A social learning approach: Vol. 4. Antisocial boys. Eugene: OR: Castalia.

Pears, K., Bruce, J., Fisher, P., & Kim, H. (2010). Indiscriminate friendliness in maltreated foster children. *Child Maltreatment, 15*, 64-75.

Pederson, D. R., Bailey, H. N., Tarabulsy, G. M., Bento, S., & Moran, G. (2014). Understanding sensitivity: Lessons learned from the legacy of Mary Ainsworth. *Attachment & Human Development, 16*, 261–270.

Pederson, D. R., Moran, G., Sitko, C., Campbell, K., Ghesquire, K. & Acton, H. (1990). Maternal sensitivity and the security of infant-mother attachment: A Q-sort study. *Child Development, 61*(6), 1974-1983.

Pettit, G. S., Bates, J. E., & Dodge, K. A. (1997). Supportive parenting, ecological context, and children’s adjustment: a seven-year longitudinal study. *Child Development, 68*, 908-923.

Pollak, S. D., Cicchetti, D., Hornung, K., & Reed, A. (2000). Recognizing emotion in faces: Developmental effects of child abuse and neglect. *Developmental Psychology, 36*(5), 679-688.

Pretzer, G. M., Lopez, L. D., Walle, E. A., Warlaumont, A. S. (2019). Infant-adult vocal interaction dynamics depend on infant vocal type, child-directedness of adult speech, and timeframe. *Infant Behavior and Development, 57,* 1-15.

Putnam, S. P., Spritz, B. L. & Stifter, C. A. (2002). Mother-child coregulation during delay of gratification at 30 months. *Infancy, 3*, 209-225.

Raby, K. L., Lawler, J. M., Shlafer, R. J., Hesemeyer, P. S., Collins, W. A., & Sroufe, L. A. (2015). The interpersonal antecedents of supportive parenting: a prospective, longitudinal study from infancy to adulthood. *Developmental Psychology, 51*, 115-123.

Raby, K. L., Roisman, G. I, Fraley, R. C., & Simpson, J. A. (2015). The enduring predictive significance of early maternal sensitivity: Social and academic competence through 32 years. *Child Development, 86*(3), 695 – 708.

Roberts, M. Y., & Kaiser, A. P. (2011). The effectiveness of parent-implemented language interventions: A meta-analysis. *American Journal of Speech-Language Pathology, 20*, 180–199.

Roggman, L. A., Cook, G. A., Innocenti, M. S., Norman, V. J., Christiansen, K. (2013). Parenting Interactions with Children: Checklist of Observations Linked to Outcomes (PICCOLO) in diverse ethnic groups. *Infant Mental Health Journal, 34*(4), 290-306.

Romeo, R. R., Leonard, J. A., Robinson, S. T., West, M. R., Mackey, A. P., Rowe, M. L., & Gabrieli, J. D. E. (2018). Beyond the 30-million-word gap: Children’s conversational exposure is associated with language-related brain function. *Psychological Science, 29*(5), 700-710.

Sanchez, M. M. (2006). The impact of early adverse care on HPA axis development: Nonhuman primate models. *Hormones and Behavior, 50*(4), 623-631.

Sanchez, M. M., Noble, P. M., Lyon, C. K., Plotsky, P. M., Davis, M., Nemeroff, C. B. & Winslow, J. T. (2005). Alterations in diurnal cortisol rhythm and acoustic startle response in nonhuman primates with adverse rearing. *Biological Psychiatry, 57*(4), 373-381.

Schindler, H. S., Fisher, P. A., & Shonkoff, J. P. (2017). From innovation to impact at scale: Lessons learned from a cluster of research-community partnerships. *Child Development, 88*(5), 1435-1446.

Sheridan, M. S., Fox, N. A., Zeanah, C. H., McLaughlin, K., & Nelson, C. A. (2012). Variation in neural development as a result of exposure to institutionalization early in childhood. *Proceedings of the National Academy of Sciences of the United States of America, 109*(32), 12927 – 12932.

Shonkoff, J. P., & Bales, S. N. (2011). Science does not speak for itself: translating child development research for the public and its policymakers. *Child Development, 82*(1), 17-32.

Shonkoff, J. P. & Phillips, D. A. (Eds.) (2000). *From neurons to neighborhoods: The science of early childhood development.* Washington, D. C.: National Academies Press.

Sidani, S., Guruge, S., Miranda, J., Ford-Gilboe, M., & Varcoe, C. (2010). Cultural adaptation and translation of measures: An integrated method. *Research in Nursing and Health, 33*(2), 133-143.

Sousa, V. D. & Rojjanasrirat, W. (2011). Translation, adaptation, and validation of instruments or scales for use in cross-cultural health care research: a clear and user-friendly guideline. *Journal of Evaluation in Clinical Practice, 17,* 268-274.

Stams, G.-J. J. M., Juffer, F., & van IJzendoorn, M. H. (2002). Maternal sensitivity, infant attachment, and temperament in early childhood predict adjustment in middle childhood: The case of adopted children and their biologically unrelated parents. *Developmental Psychology*, *38*, 806–821.

Stiles, J. & Jernigan, T. L. (2010). The basics of brain development. *Neuropsychology Review, 20*(4), 327-348.

Stolarova, M., Wolf, C., Rinker, T., & Brielmann, A. (2014). How to assess and compare inter-rater reliability, agreement, and correlation of ratings: an exemplary analysis of mother-father and parent-teacher expressive vocabulary rating pairs. *Frontiers in Psychology, 5,* 509.

Szyf, M., Weaver, I., & Meaney, M. (2007). Maternal care, the epigenome, and phenotypic differences in behavior. *Reproductive Toxicology, 24,* 9-19.

Tarullo, A., Garvin, M. C., & Gunnar, M. (2011). Atypical EEG power correlates with indiscriminately friendly behavior in internationally adopted children. *Developmental Psychology, 47*(2), 417-431.

Tamis-LeMonda, C. S., Bornstein, M. H., & Baumwell, L. (2001). Maternal responsiveness and children’s achievement of language milestones. *Child Development, 72*(3), 748-767.

Tamis-LeMonda, C. S., Kuchirko, Y., & Song, L. (2014). Why is infant language learning facilitated by parental responsiveness? *Current Directions in Psychological Science, 23*(2), 121-126.

Tottenham, N., Hare, T. A., Quinn, B. T., McCarry, T. W., Nurse, M., Gilhooly, T., … & Casey, B. J. (2010). Prolonged institutional rearing is associated with atypically large amygdala volume and emotional regulation difficulties. *Developmental Science, 13*(1), 46-61.

Tryphonopoulos, P. D., Letourneau, N., & DiTommaso, E. (2016). Caregiver-infant interaction quality: A review of observational assessment tools. *Comprehensive Child and Adolescent Nursing, 39*, 107–138.

Vernon-Feagans, L. & Bratsch-Hines, M. E. (2013). Caregiver-child verbal interactions in child care: a buffer against poor language outcomes when maternal language input is less. *Early Childhood Research Quarterly, 28*(4), 858-873.

Vilaseca, R., Rivero, M., Bersabe, R. M., Esperanza, N. P., Cantero, M. J., Ferrer, F., … Roggman, L. (2019). Spanish validation of the PICCOLO (Parenting Interactions with Children: Checklist of Observations linked to Outcomes). *Frontiers in Psychology, 10*, 1-12.

Zhang, B. (2017). Consequences of early adverse rearing experience (EARE) on development: insights from non-human primate studies. *Zoological Research, 38*(1), 7-35.

Zhou, Q., Eisenberg, N., Losoya, S. H., Fabes, R. A., Resier, M., Guthrie, I. K., … Shepard, S. A. (2002). The relations of parental warmth and positive expressiveness to children’s empathy-related responding and social functioning: a longitudinal study. *Child Development, 73*(3), 893-915.

Table 1

*Minimum Inter-Rater Reliability Thresholds*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coding Measure | % Agree  (>80%) | Cohen’s kappa  (>0.70) | ICC  (>0.80) | Pearson’s *r*  (>0.90) |
| SRS | √ | √ | √ | -- |
| WECS | √ | √ | √ | -- |
| Connection | √  (Duration) | √ | -- | √ |
| Reciprocity | √  (Duration) | √ | -- | √  (Duration) |
| CT | √  (0.5s match) | √ | -- | √  (Duration) |

*Note:* Reliability indices were selected based on the type of data coded using each instrument. The Simple Interactions subscales utilized duration-based agreement and CT reliability was calculated by analyzing coder agreement at 0.5 second intervals throughout coded film.

Table 2

*SRS Inter-Rater Reliability Across Coders During Training Week 9*

|  |  |  |  |
| --- | --- | --- | --- |
| Coder | % Agree | Kappa | ICC |
| Coder 1 | 83.13%  (SD = +/- 0.10) | 0.752  (SD = +/- 0.14) | 0.826  (SD = +/- 0.14) |
| Coder 2 | 80.63%  (SD = +/- 0.14) | 0.700  (SD = +/- 0.15) | 0.836  (SD = +/- 0.12) |
| Coder 3 | 87.50%  (SD = +/- 0.07) | 0.834  (SD = +/- 0.12) | 0.834  (SD = +/- 0.09) |
| Coder 4 | 81.25%  (SD = +/- 0.11) | 0.730  (SD = +/- 0.14) | 0.817  (SD = +/- 0.11) |
| Coder 5 | 86.25%  (SD = +/- 0.07) | 0.818  (SD = +/- 0.10) | 0.876  (SD = +/- 0.08) |
| Coder 6 | 88.75%  (SD = +/- 0.11) | 0.835  (SD = +/- 0.15) | 0.917  (SD = +/- 0.06) |
| Mean | **84.59%**  (SD = +/- 0.03) | **0.778**  (SD = +/- 0.06) | **0.858**  (SD = +/- 0.04) |

*Note*: Means are reported from pairwise comparisons of coder ratings to an expert coder. Scores are averaged across SRS training films from Week 9 of training (n=8).

Table 3

*Average SRS Reliability Across FIND ACF Films*

|  |  |  |  |
| --- | --- | --- | --- |
|  | % Agree | Kappa | ICC |
| Mean | 69.82%  SD = +/- 0.129 | 0.557  SD = +/- 0.189 | 0.779  SD = +/- 0.158 |
| Range | 43.75% - 93.75% | 0.172 – 0.913 | 0.247 – 0.95 |

*Note:* Average code reliability is based on pairwise agreement scored across all double-coded films (n=35). Large standard deviations and wide ranges across each index highlight significant variability in agreement score across films.

Table 4

*Connection Inter-Rater Reliability Across Coders During Training Week 8*

|  |  |  |  |
| --- | --- | --- | --- |
| Method | % Agree | Pearson’s *r* | Kappa |
| Duration/Sequence | 76.88%  SD = +/- .10 | 0.965  SD = +/- .05 | 0.589  SD = +/- .12 |
| Duration | 82.95%  SD = +/- .12 | 0.97  SD = +/- .04 | 0.732  SD = +/- .16 |

*Note:* Averages are reported from three coder ratings across each training film (n=8) and represent pairwise comparisons with the other two coders for each film. All reliability analyses were conducted using Noldus Observer XT coding software.

Table 5

*Average Connection Reliability Across FIND ACF Films*

|  |  |  |  |
| --- | --- | --- | --- |
| Method | % Agree | Pearson’s *r* | Kappa |
| Duration/Sequence | 79.41%  SD = +/- .09 | 0.966  SD = +/- .06 | 0.546  SD = +/- .15 |
| Duration | 83.20%  SD = +/- .10 | 0.96  SD = +/- .07 | 0.685  SD = +/- .16 |

*Note:* Averages are reported for pairwise comparisons across each of the double-coded films (n=33). All reliability analyses were conducted using Noldus Observer XT coding software.

Table 6

*Reciprocity Inter-Rater Reliability Across Coders During Training*

|  |  |  |  |
| --- | --- | --- | --- |
| Method | % Agree | Pearson’s *r* | Kappa |
| **Duration/Sequence** | 74.03%  SD = +/- .13 | 0.966  SD = +/- .05 | 0.414  SD = +/- .19 |
| **Duration** | 80.99%  SD = +/- .11 | 0.963  SD = +/- .05 | 0.638  SD = +/- .18 |

*Note:* Averages are reported across five coders and are compared pair-wise against the other four coders for each film (n=4). All reliability analyses were conducted using Noldus Observer XT coding software.

Table 7

*Average Reciprocity Reliability Across FIND ACF Films*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coding Method | Reliability Method | % Agree | Pearson’s *r* | Kappa |
| Individual | Duration/Sequence | 71.09%  SD = +/- .12 | 0.903  SD = +/- .17 | 0.427  SD = +/- .20 |
| Individual | Duration | 75.52%  SD = +/- .17 | 0.894  SD = +/- .16 | 0.603  SD = +/- .24 |
| Convergent | Duration/Sequence | 95.49%  SD = +/- .03 | 0.998  SD = +/- .003 | 0.932  SD = +/- .04 |
| Convergent | Duration | 86.00%  SD = +/- .10 | 0.949  SD = +/- .10 | 0.774  SD = +/- .14 |

*Note:* For individually coded films averages are reported for pair-wise IRR comparisons across each of the double-scored films (n=51). For convergent coded films, reliability analyses are reported for all double-scored convergent films (n=9), comparing pairwise convergent scores from 2 separate sets of raters. All reliability analyses were conducted using Noldus Observer XT coding software.

Table 8

*CT Inter-Rater Reliability Across Coders During Training*

|  |  |  |  |
| --- | --- | --- | --- |
| Tier | % Agree | Pearson’s *r* | Kappa |
| Parent Utterances | 92.82%  SD = +/- .03 | 0.975  SD = +/- .0.01 | 0.844  SD = +/- .07 |
| Child Utterances | 93.84%  SD = +/- .03 | 0.953  SD = +/- .0.02 | 0.821  SD = +/- .0.08 |

*Note:* Reliability is reported as averages across six trained coders in pairwise comparisons against a Gold Standard code. Percent agreement and Kappa were calculated by evaluating percent match on utterance determinations (yes/no) at the 0.5 second level. Pearson’s *r* calculations represent correlations between coders for utterances aggregated into 15-second bins.

Table 9

*Average CT Reliability Across FIND ACF Films*

|  |  |  |  |
| --- | --- | --- | --- |
| Tier | % Agree | Pearson’s *r* | Kappa |
| Parent Utterances | 91.75%  SD = +/- .04 | 0.898  SD = +/- .0.09 | 0.805  SD = +/- .0.09 |
| Child Utterances | 95.73%  SD = +/- .03 | 0.859  SD = +/- .0.15 | 0.719  SD = +/- .0.15 |

*Note:* Reliability is reported as pairwise comparisons across all double coded films (n=29). Percent agreement and Kappa were calculated by evaluating percent match on utterance determinations (yes/no) at the 0.5 second level. Pearson’s *r* calculations represent correlations between coders for utterances aggregated into 15-second bins.

Appendix A

The Welch Emotional Connection Scale (WECS)

A screenshot of a cell phone

Description automatically generated

Appendix B

The Serve and Return Scale Glossary

This study used an adapted version of the SRS, which contained only 3 of the original 6 subscales of the SRS measure. The 3 subscales that were used are highlighted in green within the original SRS glossary below.

**A screenshot of a cell phone

Description automatically generated**

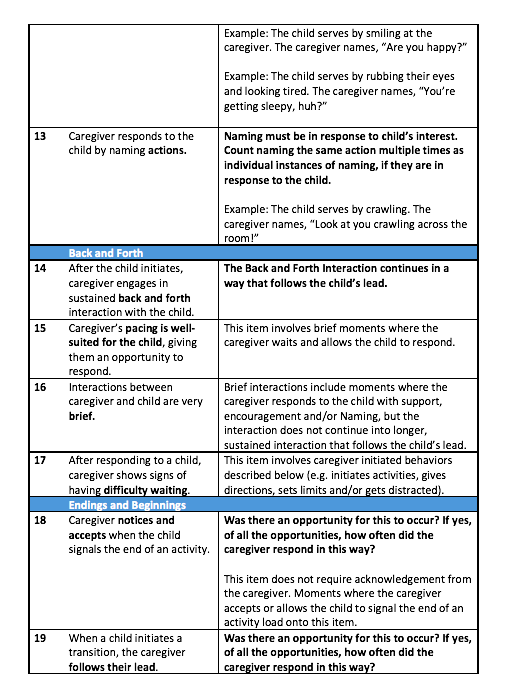
**A close up of a logo

Description automatically generated**

****

**A close up of a screen

Description automatically generated**

****

**A screen shot of a computer

Description automatically generated**

Appendix C

The Serve and Return Scale Score Sheet

A screenshot of a cell phone

Description automatically generatedThis study utilized 3 subsections from the original SRS measure. These three domains are highlighted in green below.

**A screenshot of a cell phone

Description automatically generated**

Appendix D

The Simple Interactions Glossary

**Simple Interactions (0-3) Glossary**

**\*This version of the glossary was last updated on September 30, 2019. This glossary is meant to accompany the Connection, Reciprocity, and Progression subscales for coding caregiver-child interactions with children ages 0-3.**

**Affect**: Emotion shown by facial expression and tone. Categorized broadly as positive, negative, or neutral affect.

* For something to be categorized as positive or negative affect, it must be a significant, obvious change from the neutral affective state. For example, some people seem to have half-smiles as their resting facial expression. This would not count as positive affect.

**Appropriate contribution**: Designation that happens when an activity has a leader (either parent or child) and the other participant contributes and elaborates on the activity. When the non-leading participant is a parent, he/she is considerate of the child’s pace and demonstrates patience.

* This can be thought of as a serve from the non-leader. The contribution can either be verbal or non-verbal (e.g. question, statement, or gesture) and should add new information to the interaction.
* If the child is leading, and the parent elaborates on the activity, it will only qualify as appropriate contribution if the parent also displays patience and consideration of the child’s pace. If the parent serves, but then a few seconds later interrupts the child or does not allow the child to complete a task because it is taking too long, then this would not qualify as appropriate contribution.

**Appropriateness**: A challenge is deemed appropriate when it is difficult for the child, but the child could still reasonably accomplish the task. An inappropriate challenge is above the abilities of the child at that moment.

* What matters is whether the challenge is reasonable for the specific child. No matter how much support a parent gives, a five-year-old will not be able to solve a multiplication problem.

**Challenge**: A task or situation that stretches a child’s abilities.

* There are different ways to challenge a child to do something new or difficult. From the SI tool: “Think about what is being “progressed” here: Content skill? Social skill? Behavioral skill? Character skill? It is helpful to identify that. Not all interactions focus on identifiable academic skills.”
* In most cases, asking a child to identify what something is does not count as a challenge, even if it is difficult for the child to remember. This is would just be recall. Progression involves encouraging a child to think, not just remember.

**Engagement**\*: A meaningful verbal or non-verbal contact has been established between parent and child.

* Looking at each other; one looking at the other; one looking at the same thing the other is looking at; (physical contact); (closer than combined arms’ length away **and** parent is on the same level as the child); talking.
  + \*For items in (), engagement can be ***spoiled*** if parent or child is clearly engaged with or interacting with another person or object. Spoiled engagement applies only when caregiver and child are *both* looking at/engaged with different activities (e.g. while in physical contact or close proximity).

**Fading**: The parental act of providing an appropriate challenge, offering scaffolding, but then removing that scaffolding while remaining supportive so that the child has the freedom to meet the challenge in their own way.

* From the SI tool: “It is not necessary for the child to actually MAKE progress, just that they are pushed and challenged to try to make a leap of progress (Z)”

**Interaction in joint activity**\*: Both participants are physically engaged in the same activity while close together. Proximity is required except when the activity necessitates distance (e.g., playing catch).

* Physically engaged means literally touching the activity. Resting a hand on the activity, or having a piece of the activity in hand, counts as being physically engaged.
* Even if both are doing the same activity and are very engaged, it does not count as interaction in a joint activity if they are farther than their combined arms’ length away.
* Taking turns also counts as interacting in a joint activity. For example, if the parent takes a turn for five seconds, then the child takes a turn for five seconds, then the parent takes a turn for five seconds, then this qualifies (even though there was not any time when they were both touching the activity at the same time). However, if one person takes a turn for 10 seconds, and the other person takes a turn for five seconds, then this does not qualify, because for the majority of the interval, only one person was engaging in the activity.

**Leader**\*: One participant is responsible for determining the direction of the activity through commands or by choosing a new activity.

* Evidence of leadership must happen in the first half of the interval.
* Not all commands are considered to be leadership. It is possible for an interaction to be balanced with one individual giving instructions to the other within the activity. To qualify as leadership, the command must very clearly change the course of the activity. One way to think of this is to ask yourself “is it reasonable to assume that this person might have done this without the other person asking or telling them to?”

**Looking/Talking/Touching:** Both parent and child are doing at least one of these actions (looking OR talking OR touching) with reference to the same activity.

* E.g. If the caregiver is looking at something that the child is touching, this would count as “looking/talking/touching” and should be coded CZ
* E.g. If the mom is talking about an object/activity while the child is looking at something else (completely separate), this would NOT count as “looking/talking/touching” and should be coded CY

**Mutual cooperation**: Two serve-and-return exchanges, regardless of who serves.

**Physical proximity**: Participants are within their combined arms’ reach of each other and are on the same physical level.

* If the adult is standing near the child, they are only considered to be within close physical proximity if the adult is leaning down to the child’s level.
* The adult kneeling or sitting near the child qualifies as physical proximity.
* These bullets need to be better ^ with them, maybe focus on defining what it means to be on the child’s level.
* If they are touching, they do not need to be on the same physical level.

**Return**: A basic response to a serve. A return can be verbal (e.g. answering a question, naming an object, commenting on an activity) or non-verbal (e.g. responding via a gesture, facial expression, or action).

* A question of clarification, repeating what someone said in the form of a question, or asking someone to repeat what they said, counts as a return, not a serve.
* A return does not need to be spoken. Things like nodding, head-shaking, laughing, and shrugging can all be a return.
* Looking at a person or an object (“sharing the focus”) does not count as a return. In order to count as a return, the caregiver/child must acknowledge the serve directly in their response (e.g. nodding, verbal response, smiling, gesturing).
* It is possible for a response to count as both a serve and a return. In this example, the response must acknowledge/respond to the serve directly (“return”) while also adding new information that leads to further back-and-forth interaction (the next serve). (E.g. a caregiver responds to the child’s precarious block placement by asking “What do you think would happen if it fell?”)
* If a return is split across two 15-second intervals, this should be considered a return (*not* a serve) in the second interval.

**Serve**: An initiation of communication where one person shows interest in a specific object or activity. A serve can be verbal (e.g. “what should we play with now?”; “Let’s go over there”) or nonverbal (e.g. changing gaze, gesturing, picking up an object). Any initiation of communication that adds new information is a serve.

* For younger children, repetitive movements (e.g. infant continuing to hit a toy) should be considered one serve unless there is a significant change that adds new information (e.g. response from caregiver, change in facial expression, or change in intensity of action).
  + If a repetitive action continues across intervals, rate each 15-second interval independently. For example, if a child is interacting with a toy in the same way across two intervals, this action should be counted as a new serve in the new 15-second interval.
* There must be some degree of *intention* behind an action for it to count as a serve. “Accidents” (e.g. sneezing, dropping something, falling down) are not considered serves. If a caregiver responds to an accident (e.g. saying “Bless you” or “uh oh”, reaching out to catch the child), this response is considered a serve, since (s)he is initiating communication about that activity.

**Significant communication**: At least two full serve-and-returns. Each participant must serve at least once.

**Verbal communication**: Two serve-and-returns, both from the same person; OR One serve-and-return; OR Two serves regardless of who serves.

\* For Reciprocity, this must take place for the majority of the 15-second interval.

**No Code:** Anytime video quality impacts a coder’s ability to make a decision on the flow chart, the code assigned should be “No Code.” This rule applies across a variety of contexts, including (but not limited to) the examples provided below:

* Connection:
  + E.g. Anytime one person is completely out of the frame, the code assigned should be “No Code”. (Coders cannot make a decision about “engagement” if one person is out of view).
  + E.g. If caregiver/child are engaged, but one person’s face is turned away from the camera, the code assigned should be “No Code”. (Coders cannot make a decision about matched affect).
* Reciprocity:
  + E.g. Anytime one person is completely out of the frame for the majority of a 15 second interval (i.e. >7.5 seconds), the code assigned should be “No Code.” (Coders cannot make a decision about “engagement” if one person is out of view).
  + E.g. If one person’s hands are out of view for >7.5 seconds, the code assigned should be “No Code”. (Coders cannot make a decision about whether or not there is interaction in joint activity).
  + E.g. If caregiver and child have already demonstrated significant communication (one full serve and return from each), but one person is out of frame for <7.5 seconds, the code assigned should be “RZ”. (Coders *can* make a decision about the quantity of serve and return based on the remainder of the interval).
  + E.g. If at a critical decision-making point and one person is potentially serving or returning while out of the frame, the code assigned should be “No Code”. (Coders cannot make a decision about “mutual cooperation”, “appropriate contribution”, “significant communication” or “verbal communication” if the serve or return in question happened while out of the frame).
  + E.g. If the interval is less than 15 seconds long, the code assigned should be “No Code”. (Coders cannot make valid decisions about the flow chart for intervals less than 15 seconds).

Appendix E

The Simple Interactions Connection Flowchart

**A close up of a screen

Description automatically generated**

Appendix F

The Simple Interactions Reciprocity Flowchart

**A close up of a screen of a cell phone

Description automatically generated**

Appendix G

The Conversational Turns Scale

*The Conversational Turns (CT) Scale is meant to be a simple, straightforward coding scheme that can be applied across a variety of contexts to quantify caregiver and child language. The coding rules below serve as guidelines for what counts as an utterance, where utterances start and stop, and for whom utterances should be coded. This coding scheme was originally designed for use with ELAN Linguistic Annotator software.*

a) Mark any sound that you think was made by the vocal tract (i.e. larynx, throat, mouth, lips, etc.). In order to count as an utterance, the sound must have communicative intent (e.g. words, sound effects, gurgling). Coughing, sneezing, and involuntary sounds from the vocal tract should not be marked as utterances.

b) Be as accurate as possible when setting utterance boundaries – try to be within tenths of a second. Pauses between phrases that are greater than 0.25 seconds and/or obvious changes in sound effects to indicate a change in concept should be marked as separate utterances.

c) Caregiver and child utterances should ONLY be coded for the target child and target caregiver in the dyadic interaction. Do not mark utterances from other adults, children, or electronics (including toys, radio, music, or TV).