Measuring the Quality of Caregiver-Child Interactions for Infants and Toddlers (Q-CCIIT)

January 2015
OPRE Report 2015-13

Q-CCIIT

This page has been left blank for double-sided copying.

FINAL REPORT

Measuring the Quality of Caregiver-Child Interactions for Infants and Toddlers (Q-CCIIT)

OPRE Report 2015-13

January 2015

Sally Atkins-Burnett
Shannon Monahan
Louisa Tarullo
Yange Xue
Elizabeth Cavadel
Lizabeth Malone
Lauren Akers

Submitted to:
Office of Planning, Research and Evaluation
Administration for Children and Families
U.S. Department of Health and Human Services

Project Officer: Amy Madigan
Contract Number: HHSP23320095642WC/HHSP23337016T

Submitted by:
Mathematica Policy Research
1100 1st Street, NE
12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763

Project Director: Louisa Tarullo
Reference Number: 06861.800

Suggested citation:

Atkins-Burnett, Sally, Shannon Monahan, Louisa Tarullo, Yange Xue, Elizabeth Cavadel, Lizabeth Malone, and Lauren Akers (2015). *Measuring the Quality of Caregiver-Child Interactions for Infants and Toddlers (Q-CCIIT).* OPRE Report 2015-13. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U. S. Department of Health and Human Services.

This report and other reports sponsored by the Office of Planning, Research and Evaluation are available at http://www.acf.hhs.gov/programs/opre/index.html.

OPRE

MATHEMATICA Policy Research

DISCLAIMER

The views expressed in this publication do not necessarily reflect the views or policies of the Office of Planning, Research and Evaluation, the Administration for Children and Families, or the U. S. Department of Health and Human Services.
This report and other reports sponsored by the Office of Planning, Research and Evaluation are available at http://www.acf.hhs.gov/programs/opre/index.html.

ACKNOWLEDGMENTS

The authors would like to express their appreciation to our project officer Amy Madigan and other federal staff at OPRE and the Office of Head Start. We thank the Mathematica team including Eileen Bandel, Veronica Barrios, Sara Bernstein, Timothy Bruursema, Felicia Hurwitz, Ashley Kopack Klein, Marisa Putnam, Kimberly Ruffin, Jillian Stein, Rebecca DiGiuseppe, Kevin Manbodh, Jessica De Santis, Jessy Nazario, Season Bedell, Irene Crawley, Judy Cannon, Sameena Salvucci, Amang Sukasih, and Xiaojing Lin. We appreciate the efforts of our able cadre of field staff. We thank our partners, Tamara Halle,Kathryn Tout, Rachel Anderson and Amy Blasberg (Child Trends), who conducted the initial literature review, Margaret Burchinal (University of North Carolina-Chapel Hill), who consulted on analysis, and Kerry Kriener-Althen and Gabriela Lopez (WestEd) who drafted the sustainability plan. We are grateful for the contributions of our Technical Work Group, including Robert Bradley, Judy Carta, Martha Edwards, Karen Heying, Judith Jerald, Ron Lally, Tammy Mann, Lori Roggman, Susan Sandall, Kathy Thornburg, and Deborah Vandell. We thank our original project officer, Rachel Chazan Cohen, for her support. Most of all, we offer our thanks to the staff, families, and children of the child care programs across the country, who opened their doors and shared their time with us.

CONTENTS

APPENDICES

This page has been left blank for double-sided copying.

TABLES

FIGURES

This page has been left blank for double-sided copying.

OVERVIEW

The Quality of Caregiver–Child Interactions for Infants and Toddlers (Q-CCIIT) observation tool was developed to measure the quality of child care settings—specifically, the quality of caregiver-child interaction for infants and toddlers in nonparental care. This tool is appropriate for use across child care settings, including center-based care and family child care homes (FCCs), as well as single- and mixed-age classrooms. This tool offers early childhood professionals and researchers the means to obtain a better understanding of how caregivers and young children interact in child care settings and improve child care services in the future.

The Q-CCIIT observation tool measures caregiver support for social-emotional development, cognitive development, and language and literacy development as well as areas of concern. Specifically, the Q-CCIIT requires observation of 10-minute time samples [1] that capture interactions occurring with a given caregiver and child/group of children at a given time, as well as global ratings based on the entire observation time. The Q-CCIIT allows observers to code some dimensions within each 10-minute sample (a cycle) while coding other dimensions across the entire observation period. During each cycle, observers use the Q-CCIIT rating form to take notes that provide the evidence for both the cycle ratings and the dimensions rated across the visit at the end of the observation. The caregiver is rated based on the average experience provided to the children in each cycle. When multiple caregivers are present in a setting, the observer focuses on a different caregiver in each cycle, requiring an observation lasting a minimum of 2 hours.

We used a four-phase approach to develop, operationalize, and refine the Q-CCIIT measure and collect data on its psychometric properties: an initial phase, comprising a literature review and the development of a measurement framework, and three data collection phases we refer to as the pretest, pilot test, and psychometric field test. The number of observations, geographic locations, and observers increased with each phase of data collection. With each phase, we refined the measure until we ultimately evaluated the psychometric properties of the final measure during the field test. The final field test sample included 400 classrooms (110 FCCs and 290 center-based classrooms) in 10 geographical clusters spanning 14 states and the District of Columbia.

These field test analyses provide psychometric evidence supporting the reliability and validity of the Q-CCIIT as a measure of caregiving quality. Adequate to strong reliability was found across multiple analytic methods. The Generalizability study (G-study) indicated that most of the variance may be attributed to differences in classrooms and the interaction of classrooms with items and cycles. Decision study (D-study) results indicated that the G-coefficient and dependability index (phi) for each of the three support areas showed a good level of reliability. Both confirmatory factor analysis (CFA) and item response theory (IRT) Rasch analyses supported the construct validity of the Q-CCIIT. Differential item functioning (DIF) analyses generally suggested comparable item difficulties for the Q-CCIIT scales by child age, program type, and concentration of dual language learners (DLL). Convergent validity with related measures was evident, with expected moderate to high-moderate relationships found with the Observational Record of the Caregiving Environment (ORCE) in all settings and with the Infant/Toddler Environment Rating Scale-Revised (ITERS-R) or Family Child Care Environment Rating Scale-Revised (FCCERS-R) depending on setting type. We also found some evidence for discriminant validity of the

[1] In the psychometric field test, we observed six 10-minute time samples in each classroom and family child care home.

scales. Caregiver characteristics had a weak relationship with Q-CCIIT scales, and child/adult ratios were not related to any of the Q-CCIIT scales.

Although only psychometric work has been done on the Q-CCIIT, the measure's strong reliability, sensitivity to variation in caregiving, and evidence of validity support its ability to provide estimates of quality across and within caregivers and suggest its utility for the potential uses of professional development, evaluation, and research. The Q-CCIIT offers the opportunity to identify strengths and challenges in caregiving in a variety of settings and the potential to test different approaches for improving caregiving for children.

I. QUALITY OF CAREGIVER-CHILD INTERACTIONS FOR INFANTS AND TODDLERS (Q-CCIIT) OBSERVATION TOOL

With 40 percent of infants and 50 percent of toddlers receiving nonparental care on a weekly basis (Mulligan et al. 2005), the ability of child care providers to develop nurturing relationships with the young children in their care is central to the development of young children nationwide. Interactions with caregivers are the active ingredients through which relationships form and children's early communication, learning, and competence unfold (Shonkoff and Phillips 2000). The role of interactions and early relationships in children's emerging abilities extends beyond experiences encountered in the home. Extensive reviews of research conclude that the relationship between young children and their nonparental caregivers is an important contributor to children's well-being (Shonkoff and Phillips 2000; Shonkoff 2003; National Research Council 2003).

Measures appropriate for assessing the quality of interactions between caregivers and infants and toddlers in child care settings are scarce. Currently available measures often focus more on the environment and/or are for limited use—for example, those valid only for certain types of settings. The Office of Planning, Research and Evaluation (OPRE) in the Administration for Children and Families (ACF) of the U.S. Department of Health and Human Services contracted with Mathematica Policy Research to develop a research-based measure valid for use with infant and toddler classrooms as well as center-based and family child care (FCC) settings. After more than 50 years of public debate about whether infants and toddlers should be in child care, discussion has shifted to how best to provide nurturing, safe child care environments for young children. Central to this question is how we define and measure the quality of child care. With the rise in the use of observational measures of child care in states' monitoring systems, the availability of reliable and valid measures of infant/toddler care is critical for policymakers, program developers, child care providers, parents, and researchers. Another factor driving demand is the need for tools that can support the professional development of caregivers in these settings. Few measures are appropriate for assessing the quality of infant/toddler care settings, and those that do exist have limitations, ranging from being too specific to a certain age or type of setting to a lack of a solid record of reliability and validity (Zaslow et al. 2011). Some measures currently available were developed as downward extensions of observations for older preschool children's settings, which differ in structure and content from those for younger children.

The Quality of Caregiver-Child Interactions for Infants and Toddlers (Q-CCIIT) observation tool is designed to assess the quality of child care settings—specifically, the quality of caregiver-child interactions for infants and toddlers in nonparental care. This tool offers early childhood professionals and researchers the means to obtain a better understanding of how caregivers and young children interact in child care settings and improve child care services in the future.

Our goal in assessing the quality of caregiver-child interactions is to evaluate the dimensions of caregiver interactions that lead to more positive outcomes for infants and toddlers. In the process of developing the Q-CCIIT observation tool, we examined the available research evidence for different behaviors and constructs; the interactions between caregivers and young children are at the core of this observation tool. In this chapter, we describe our approach to selecting the key constructs that guided measurement development. In selecting them, we drew on research and, where research was limited, recommended best practices. Very little research on

infants and toddlers focused on interactions in group care settings. In developing the Q-CCIIT observation tool, we kept in mind the multiple purposes that Q-CCIIT may address—monitoring, accountability, and professional development.

A. The Q-CCIIT conceptual model

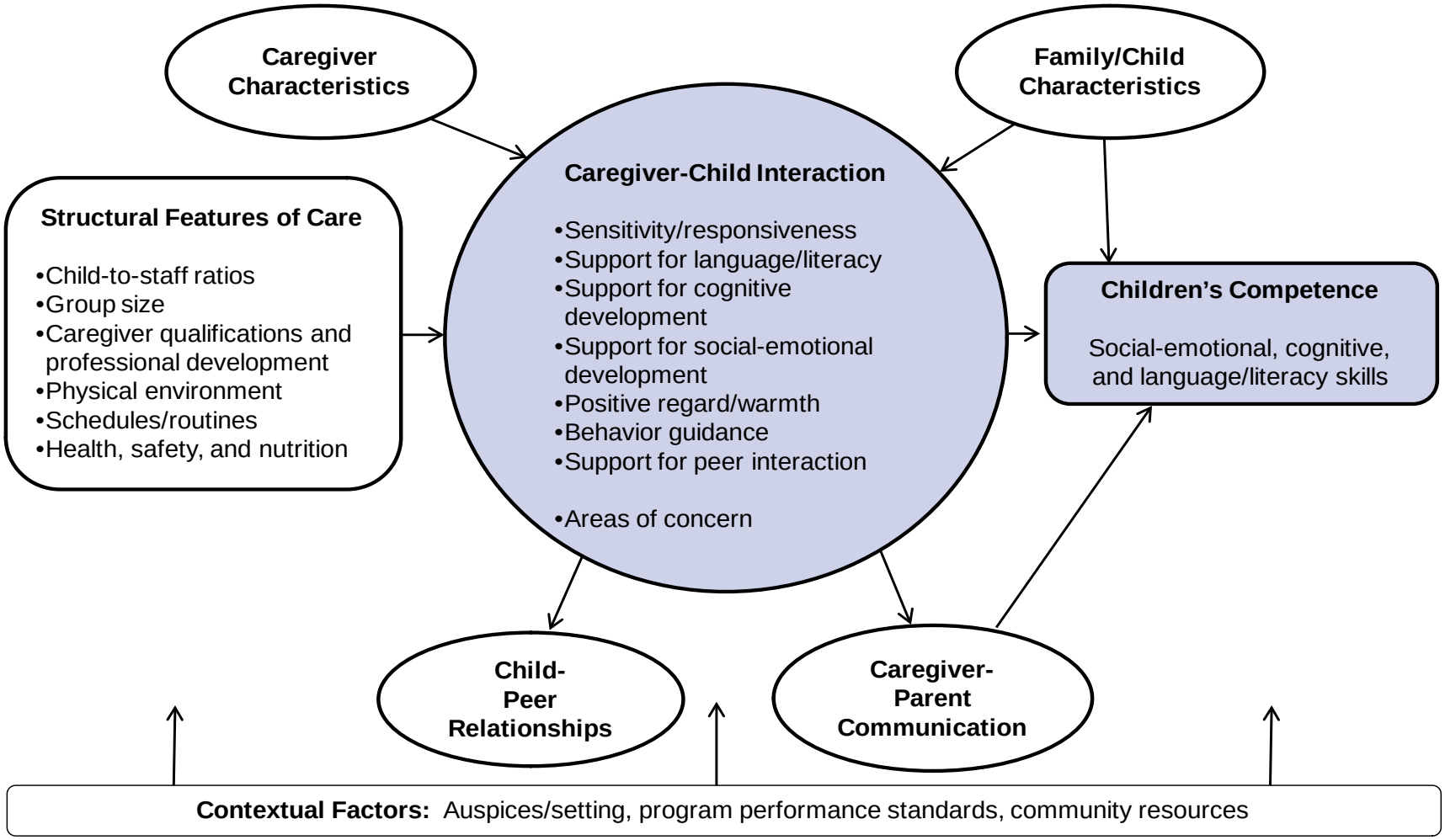
The Q-CCIIT observation tool was developed to measure the quality of child care settings—specifically, the quality of caregiver-child interaction for infants and toddlers in nonparental care. This tool is appropriate for use across child care settings, including center-based care and FCC homes, as well as single- and mixed-age classrooms.[2]

Drawing on the research on early development of infants and toddlers, we created a conceptual model to represent the areas to assess in a measure of caregiver interaction with very young children, also including those factors that influence caregiver interactions, such as the activities and safety and organization of the environment. As illustrated in the conceptual model (Figure I.1), indicators of process quality—such as sensitive and interesting caregiver-child interactions—directly influence children's social-emotional, cognitive, and language development. Additionally, structural features of care, such as caregiver professional development and child-to-staff ratio, affect children's development indirectly, through influences on process quality (NICHD Early Child Care Research Network 2002a; National Research Council 2003; Phillipsen et al. 1997; Vandell and Wolfe 2000). The quality of caregiving interactions is also influenced by the communication between caregivers and parents; children's interactions among peers; and the characteristics of the family, child, and caregiver. Family/child characteristics and caregiver-parent communication can also be direct influences on children's development. Relationships are transactional, and individuals' own behavioral styles contribute to and shape the quality of interactions with others (Sameroff 2009; Sameroff and Chandler 1975).

The conceptual model incorporates the broad range of factors that influence the quality of care that infants and toddlers receive. The Q-CCIIT observation tool primarily focuses on the indicators of positive caregiver-child interaction in the large shaded circle while recognizing other influences on the relationship between quality and child competence.

[2] We use the term "classrooms" throughout this document to signify groupings in center-based and FCC settings.

Figure I.1. Research-based conceptual model for infant/toddler quality of care



B. Selection of constructs

We used the conceptual model and a review of the research on parent-child interactions with infants and toddlers (Halle et al. 2011) to guide the selection of constructs. We also explored the literature on peer interactions and practices associated with the cognitive and language development of infants and toddlers, paying special attention to practices in group settings.

The review of parent-child interaction research and measures indicated that the areas of focus included responsiveness, sensitivity, affect, empathy, warmth, touch, language, joint attention, praise, verbalization, synchrony, directiveness, and intrusiveness (Halle et al. 2011). These constructs have been conceptualized and defined in various ways across research studies. For example, sensitivity is sometimes defined as akin to parental warmth and acceptance (De Wolff and Ijzendoorn 1997); alternatively, sometimes it is viewed as more similar to what some researchers call responsiveness (Martin et al. 2007; Ryan et al. 2006; Tamis-LeMonda et al. 2004). In selecting and defining our constructs, we also looked at negative aspects of caregiving separately from positive features, since the two could co-occur. Thus, we defined unique dimensions rather than making negative and positive behaviors two extremes of the same dimension. For example, caregivers may build positive relationships with children by using nurturing touch, smiling, and speaking children's names as they comment positively, while also occasionally being physically harsh by spanking or pulling a child. In high quality settings, we want to see high levels of positive caregiving behaviors and the absence of negative behaviors. Knowing about both positive and negative behaviors provides information that is important for understanding what is needed to improve the quality of care.

The Q-CCIIT includes the constructs with the most consistent research evidence. Differences in how researchers defined constructs affected our efforts to select and operationalize them. Some researchers defined and examined constructs individually while others combined different constructs based on theory (for example, Dodici et al. 2003) or empirical analysis. At times, researchers used the same measure (for example, Infant/Toddler Home Observation for Measurement of the Environment (IT-HOME) and the National Institute of Child Health and Development [NICHD] Three-Bag task) and individual constructs and indicators but analyzed them in multiple ways, resulting in slightly different factor structures when examined with different groups (Halle et al. 2010, 2011). To date, the analyses required to determine the unique contribution of different constructs in the conceptual model to the association with child outcomes have not been available. With that in mind, we included the parent-child interaction constructs represented most consistently in factors correlated with child outcomes. Similarly, we included constructs that research found to be associated with cognitive and language development for infants and toddlers. In some cases in which the research base is weak (particularly related to support for peer social play and cognitive development in these early years), we included constructs reflecting expert-recommended best practices.

C. Roadmap

In this chapter, we describe the conceptual framework that formed the basis for the development of the Q-CCIIT observation tool. In Chapter II, we describe the dimensions measured by Q-CCIIT. In subsequent chapters, we detail the development, administration, scoring and potential uses, and technical properties. In Chapter III, we outline the observation methods and describe the Q-CCIIT tool itself—the items and the observation rubric and scoring sheets. We discuss in Chapter IV how to score, interpret, and use the Q-CCIIT results. In

Chapter V, we describe the development process. We describe the sample and methods used in the psychometric field test in Chapter VI. In Chapter VII, we focus on the Q-CCIIT tool's technical properties—reliability and validity data from the psychometric field test.

This page has been left blank for double-sided copying.

II. DIMENSIONS OF CAREGIVER-CHILD INTERACTIONS SUPPORTING DEVELOPMENT

In this chapter, we discuss the constructs measured by the Q-CCIIT including those that support social-emotional, cognitive, and language development, and the practices needed to provide a supportive environment for responsive caregiving in group settings. Sensitive caregiving and responsiveness are related to all aspects of development and are a necessary condition for high quality care. Aspects of responsive caregiving are woven into the higher ratings of most of the items. From a theoretical perspective, how the caregiver sensitively orchestrates different practices, behaviors, and dispositions within an interaction with particular children is what leads to quality.

In the sections that follow, we describe the constructs included within each domain of caregiving practice. When available, we report the empirical and theoretical foundation for the constructs and key dimensions of quality caregiving included in each domain. As previously noted, when empirical evidence was not available, we drew on recommended best practices for some constructs, particularly those related to balancing and managing the care of multiple children in group settings.

A. Supporting social-emotional development

Responsive caregiving, also referred to as responsiveness or sensitive caregiving, involves being sensitive to the child's cues and responding contingently to them. Child care and parent-child relationship research has consistently demonstrated that responsiveness is related to positive child outcomes (Bornstein and Tamis-LeMonda 1989; Mahoney et al. 1996; Poehlmann and Fiese 2001; Ryan et al. 2006; Tamis-LeMonda et al. 2004; Zeanah 1993).

At the most basic level, caregivers must be emotionally available and sensitive to the needs and interests of the child. Responsiveness includes both physical and verbal responsiveness to the child (both during times of distress and not), and requires responding to the child's engagement and disengagement cues (Kelly and Barnard 2000). In parenting situations, maternal responsiveness is defined as a direct response to an infant's behavior and an action that relates to the behavior "conceptually and temporally" (Bornstein and Tamis-LeMonda 1989). Caregivers can be responding to distress signals (such as crying); non-distress signals, such as bids for attention, vocalizations, or play; or child disengagement or other emotional cues. Young children also need support in understanding and regulating strong emotion. Caregivers provide support in a variety of ways by providing words for feelings and demonstrating strategies for dealing with feelings (Joseph et al. 2011; Ostrosky et al. 2011b).

Frequent positive and responsive interactions help caregivers establish supportive, stable relationships with children. One core piece of such relationships is warmth, which typically is defined as including aspects of positive affect or positive regard[3] and nurturing touch. Positive affect/positive regard refers to the ways the caregiver communicates positive feelings for and with the children, which supports children's own behavior—both positive and negative—toward

[3] At times, the observation of warmth may include the absence of negative affect, negative regard, or detached or flat affect. We focus here on the presence of positive features that support development. Negative aspects of caregiving are discussed further in Section D.

others. Positive regard or affect, as a demonstration of warmth, can take a variety of forms—smiling, praise, positive verbal comments, facial expression, and tone, for example. Parental warmth and positive affect have been shown to have a relationship with young children's engagement, mutuality, and positive affect during interactions, as well as more broadly for social functioning (Forbes et al. 2004; Fuligni et al. 2004; Ispa et al. 2004; Steelman et al. 2002). Although not examined in nonparental care settings, greater demonstration of warmth by parents correlates with less child aggressive behavior or negativity during toddlerhood (Fuligni et al. 2004; Ispa et al. 2004).

The level of positive regard or affect shown often differs across ethnic groups. This may be due to the sensitivity of raters to the different levels of effusiveness demonstrated by caregivers from different cultural backgrounds (Wang et al. 2007). In some cultures, communication of positive feelings may be subtle, and infants likely learn the caregiver's signals. In group care, however, children may come from diverse backgrounds; thus, a broad set of behaviors should be considered. For example, smiling (not smirking) is universally interpreted as a positive and welcoming signal. Smiling typically elicits smiling. Not only do caregivers who smile at children elicit more smiling from them, but children who smile more often elicit more interaction from caregivers and usually have more friends as they grow older (Meisels et al. 1996). Thus, a measurement of warmth and positive affect should include caregivers' frequently smiling at all children—including children who do not smile much.

Some measures of caregiver-child interactions also include use of praise as an indicator of positive regard. Praise could provide one example (but not the only one) of ways that caregivers build a positive relationship with a child. It may also be tied to language development, as Hart and Risley (1995) found that parents who talked more frequently also provided more praise and positive commenting.[4] Thus, the use of positive commenting within parallel language (that is, a caregiver's descriptions of what children are doing) provides one avenue for assessing praise to support development.

In measuring caregiver warmth, the importance of nurturing touch for positive social-emotional development of children in group care has been established. Studies of children in orphanages and other institutions, as well as work with primates, clearly illustrate the importance of nurturing touch beyond task-oriented or functional caregiving touch (Field 2001; Montagu 1971). In recent years, studies of positive touch show a relationship of touch to physical development, emotional development, regulation, and brain development (Carlson and Nelson 2006; Drehobl and Fuhr 2000; Field et al. 1996; Field et al. 1997; Hernandez-Reif et al. 2001; Kaler and Freeman 1994; Perry and Pollard 1997; Perry and Szalavitz 2007; Schneider 2006). However, the measurement of touch must be considered carefully, so as to be sensitive to cultural taboos (for example, touching or stroking the head in some Southeast Asian cultures is taboo). Common examples of positive touch across cultures include hugs, holding, and carrying (particularly for non-mobile infants).

Our original conceptualization of support for social-emotional development included building positive peer relationships as well as adult-child relationships. With the exception of the work of Howes and colleagues (Howes 1980, 1988; Howes and Matheson 1992; Howes et al.

[4] The Hart and Risley sample was not culturally diverse. Praise is viewed differently across cultures and is taboo for young children in some Asian cultures (Bernstein et al. 2005; Chan and Lee 2004).

1989), most of the work on peer interaction examines preschoolers or older children. However, infants and toddlers have even greater need for caregiver support to facilitate positive social integration than do preschoolers (Kryzer et al. 2007). Creating opportunities for peer interaction is an important ingredient for the development of positive social skills (Ladd et al. 1988; Ladd 1992) and encouraging friendship formation and maintenance appears to be more important than simply providing time for interacting with peers (Howes 1983; Ladd and Hart 1992; Parke et al. 1992).

Our empirical results indicate that items assessing support for peer interaction were observed infrequently and with lower quality than other areas of support for social and emotional development. The peer interaction items include a cognitive component (recognizing that peers are people and learning how to problem solve) and were associated more strongly with the items assessing support for cognitive development; we discuss these in Section B.

In addition to providing support for specific areas of children's development, caregivers demonstrate behaviors—such as joint attention, routines, and limits—that support children's development across domains. In supporting development, joint attention refers to the caregiver's and child's shared attention to an object or event and may be measured as the amount of time that the parent and child look at or share attention toward an object (Dodici et al. 2003) or the number of joint attention episodes (Markus et al. 2000). Both the number of joint attention episodes and the length of time engaged in joint attention have been related positively to language development (Dodici et al. 2003; Markus et al. 2000). While joint attention supports development across domains, we focus on defining it for its support of particular developmental domains. We capture joint attention in a variety of ways across domains such as shared attention to the child's social bids to support social-emotional development, shared attention during object exploration to support cognitive development, or shared attention demonstrated through parallel language and book sharing that supports language development.

Routines are one way in which caregivers provide consistency in the child's life and allow the infant or toddler to anticipate events and interactions (Cook et al. 2004; Winton et al. 2008). Clear research evidence is not available, but use of routines long has been recognized as a best practice with young children (Bovey and Strain 2011; Ostrosky et al. 2001a). Some routines (such as schedules for eating and sleeping) have to be more flexible for young infants as they learn to regulate physiological processes. Overall, the use of routines helps to support children's self-regulation.

Positive limit setting and consistent behavioral guidelines also are important for helping older infants and toddlers to regulate behavior (Cook et al. 2004). The use of positively worded limits and prompts (including anticipatory prompts), along with consistency in enforcing behavioral guidelines, helps children understand and meet behavioral expectations. Measuring the occurrence of this behavior provides information on caregiver support, as very young children have difficulty in comprehending negative statements and often respond to the end of a statement (for example, "Stop jumping!" may lead to more jumping). In addition to supporting individual children, the management of the classroom helps children manage their own behavior.

1. Dimensions of caregiving supporting social-emotional development

The dimensions identified by the research as essential to assessing responsive caregiving in support of infant and toddler social-emotional development are the following; all of these are measured by the Q-CCIIT:

- **Responding Contingently to Distress.** Distress includes crying, having a tantrum, or otherwise showing signs of acute distress. Key features of this dimension include the immediacy of a caregiver's response (or at least acknowledgment) as well as matching that response to the child's level of distress.
- **Responding Contingently to Social Cues.** Social cues include verbal and non-verbal requests for attention or mutual enjoyment, the child's vocalizations, or bids for play. The caregiver's immediacy and individualization in responding to children determine the quality of the interaction.
- **Responding to Emotional Cues.** Caregivers who appropriately respond to emotional cues demonstrate that they are emotionally available to children and sensitive to their emotions (and anticipate the need for regulatory support). They share children's enjoyment and other positive emotions. They notice when children are unhappy or disengaged and try to comfort or involve them in activities. They may support regulation by commenting on the child's emotional state and offering strategies for dealing with emotions. They may also help children regulate positive emotions that would otherwise overwhelm the child.
- **Building a Warm, Positive Relationship.** Caregivers may use a variety of ways to build a warm, positive relationship, such as smiling, using children's names in positive ways, making positive comments, showing interest in their activities, sitting in close proximity to the children, or supporting children's attempts at autonomy. Positive regard and warmth help to build positive relationships between caregivers and children. When the relationship is warm and positive, children will often seek out the caregiver, not only for help, but to share positive experiences, and caregivers respond warmly to these bids.
- **Supervising Play and Activities.** Supervision of play provides information on how involved a caregiver is in children's play and how much she or he models, supports, or shows interest in children's play. When caregivers pay attention to and join in a child's play (without taking over), they communicate that children are safe and what they do is interesting and important to the caregiver.
- **Responsive Routines.** Caregiver use of responsive routines and schedules supports children in self-regulation and allows them to anticipate what will happen next. When children have a routine and know what will happen throughout the day, they are able to relax and explore their environment. It is important to capture evidence of routines—both visual and verbal—and the individualization of routines for children's developmental levels and needs.
- **Classroom Limits and Management.** Clear limits are important for children's sense of predictability and comfort as well as overall classroom management and safety. Caregivers demonstrate this dimension by using positive classroom limits, showing appropriate behaviors, and supporting children's positive behaviors.

- **Sense of Belonging.** A sense of belonging provides an indicator of the extent to which the caregiver has created a warm, welcoming environment in which children are acknowledged and see themselves as valued members of the community, rather than outsiders.

B. Supporting cognitive development

Supporting cognitive development includes providing materials and experiences that allow children to develop an understanding of the world and how it works, and helping them to organize their knowledge (using concept development) to reason and solve problems. Caregivers may also actively expand children's skills and abilities through explicit teaching (calling children's attention to specific knowledge and concepts or intentionally providing experiences that extend children's knowledge beyond everyday experiences), and by promoting and guiding children's exploration of diverse and stimulating environments (Martin et al. 2007). Cognitive support includes support for object exploration and concept development; scaffolding for higher levels of play; and helping children to reason (learn about cause and effect, spatial relations, and problem solving), understand different perspectives, and develop a theory of mind and memory skills, along with broadening their knowledge of the world.

From birth to three years, children's understanding of the world grows rapidly, and caregivers may focus on different skills and knowledge as children develop. Research on interactions between mother and child notes differences in the cognitive support provided to children across this time period. For example, early maternal responsiveness to very young infants includes affectionate touch and "motherese" vocalization in response to the infant's level of alertness; at 6 months and beyond, joint attention and the joint exploration of novel objects is associated with cognitive development (Feldman et al. 2002, 2004; Landry 1986; Ruff 1986). Between 9 and 21 months of age, maternal responses to children's exploration of objects decrease, while responses to play prompts increase (Tamis-LeMonda et al. 2001).

Including object exploration in the measure provides an opportunity to understand early cognitive support. More active exploration of objects is associated with earlier understanding of cause and effect in infants (Lobo and Galloway 2008). Adults can support object exploration by positioning young infants (for example, putting them into a sitting position so their hands are free to explore) as well as directly helping infants to explore objects. Using randomized controlled trials, Lobo and Galloway (2008) found that changes in postural support and guided experience with objects led to earlier exploration of objects. As children develop fine and gross motor skills, they are able to explore their world more independently, increasing their understanding of cause-and-effect relationships, exploring attributes of objects (including object permanence), solving problems, and learning new ways to act on their environment. Caregivers who provide opportunities for exploration and learning not only introduce new objects and experiences, but also show children new ways to use objects by directing attention to novel things in the environment and relationships between actions and events, repeating as needed and encouraging child imitation.

As infants and toddlers explore objects, adults may be emotionally available and sensitive to the child in play and yet not foster higher levels of play. When caregivers scaffold higher levels of play, they model and expand the child's play with objects—illustrating attributes of objects by rolling round objects, for example, or suggesting and demonstrating different ways to manipulate and use materials, and later helping toddlers to move from functional to representational play.

Caregivers may scaffold play in a variety of ways. Didactic approaches to play include structuring play through using props and verbal prompts, establishing and arranging the play context, guiding children in taking on and maintaining roles or actions, demonstrating, engaging as a play partner, and providing feedback (Fenson and Ramsey 1980; O'Reilly and Bornstein 1993). Toddlers' play with an adult usually includes more symbolic play than when they play alone (O'Reilly and Bornstein 1993; O'Connell and Bretherton 1984), and play in which the caregiver provides prompts is often more sophisticated than spontaneous play with a caregiver or play alone (O'Reilly and Bornstein 1993; Tamis-LeMonda and Bornstein 1991).

Caregivers also foster children's cognitive growth through developing their knowledge of concepts, such as size, shapes, color, and spatial concepts (over/under, between, up/down). Relationships between concept development and child outcomes have been found in studies using the IT-HOME and the Child Care Home Observation for Measurement of the Environment (CC-HOME) (Bradley et al. 2003; Fuligni et al. 2004). The HOME (Caldwell and Bradley 1984) examines the number of pre-academic concepts (letters, colors, shapes, and numbers) introduced by the caregiver. Attention also should be paid to whether caregivers introduce and talk about such concepts as feelings, attributes of objects or actions, spatial relations, and categories.

Peer interaction activities foster cognitive growth. How caregivers foster positive peer relationships in early childhood settings that include multiple children—those who are of the same age and of different ages—is an area of interest for researchers. In learning peer interaction, infants must first learn that peers are people rather than toys, and then learn how to negotiate interactions with peers who are not as sophisticated in communication as adults. As children grow, their interactions also include solving problems with peers, negotiating roles and sharing materials, and understanding the point of view of their peers.

Peer play for infants and toddlers usually involves toys, evolving from reciprocal play and peer imitation among 1-year-olds to social pretend play among 2-year-olds (Howes 1988). Caregivers support social play and peer interaction by ensuring fair distribution of materials; encouraging early peer imitation; commenting positively on shared interests among children; providing play opportunities that pair children with similar interests; and modeling and encouraging peer interaction for sharing, empathy, gentle behavior, and positive ways of solving social problems (Cook et al. 2004; Guralnick et al. 2003; Ladd 1992; Parke et al. 1992).

With the exception of the introduction of concepts, the explicitness of instruction often is examined in the research literature only as to directiveness. The influence of directiveness is unclear: it has been related both positively and negatively to child outcomes. Landry and colleagues (1997) found differences in the association by age: directiveness was positively related to younger children's growth (found at 18 months of age, but not at 40 months). It may be that the level of directiveness needs to be balanced against the child's ability to provide cues to the caregiver or engage in independent play. Marfo et al. (1998) hypothesized that parents of children who are less responsive may adapt to a paucity of child cues by becoming more directive. However, differences also have been found among cultural groups as to the strength and presence of a relationship between directiveness and child outcomes (Johnston and Wong 2002). Directiveness is defined in different ways across studies, and it may be that a distinction should be made between more explicit instruction in support of child interests and directiveness that is intrusive and does not respond appropriately to the child's cues. In measurement,

developers should take care to define and provide examples of intentional and explicit teaching responsive to children's interests.

1. Dimensions of caregiving supporting cognitive development

The dimensions indicated by the research as important for caregiver support for infant and toddler cognitive development are as follows. All are measured by the Q-CCIIT:

- **Support for Object Exploration.** Caregivers' support will vary by the child's age. Supporting infant object exploration includes positioning or providing objects; support for toddler object exploration expands to modeling (learning to imitate new movements) and verbal support for trying new actions and learning about the qualities and features of objects (for example, how to make things move or open; turning items to make them fit in a space; sorting objects with and without wheels). Additional strategies include caregiver's scaffolding of children's exploration by helping to position objects and adding materials that expand exploration (for example, adding a train track or ramp to expand the opportunities for a child rolling a car).
- **Scaffolding Problem Solving.** Problem solving provides children with a way to understand means-end relationships and frequently encountered problems of everyday life. Caregivers may demonstrate support through the prompts provided to children, explicit discussion of the problem, and more direct scaffolding.
- **Extending Representational Play.** Play begins with sensory-motor play and advances to representational (pretend) play that involves symbolic use of materials or pretend actions and roles. To support development, caregivers should demonstrate the ability to structure, reinforce, and extend play to higher levels.
- **Support for Concept Development.** Caregivers' interactions with children may support concept development by the language (for example, labeling of objects, categories, or feeling) and activities provided. Common concept categories include descriptors/attributes (for example, color, shape, size), spatial relations (for example, under/above, up/down, and in/out), pre-academic (for example, letters and numbers), and emotions. Capturing the differences and extent of concept development by caregivers across the day (by means of a checklist) can provide a picture of exposure to cognitive concepts.
- **Giving Choices.** Providing realistic choices and scaffolding how children make choices can help them develop important cognitive skills (such as weighing alternatives and understanding that in making a choice, another option is no longer available). For infants and toddlers, limiting choices and helping with selection can facilitate success. Choices may not be available in every activity; for example, in some family cultures, children are expected to be grateful for any food they receive and not given a choice of food or drink. Differentiating when they can and cannot make choices is another learning task for this age group. Caregivers offer support for understanding when and how to make choices, using verbal and visual prompts.
- **Explicit Teaching.** This construct captures intentional efforts to expand the child's knowledge of the world. The ways in which caregivers organize experiences to increase children's knowledge of the world may vary. In measuring explicit teaching, indicators range from providing exposure to ideas or information to directly instructing children in new

concepts or actions, and organizing the information in ways that help children make connections.

- **Supporting Peer Interaction and Play.** This construct examines the variety of strategies caregivers use to encourage and support peer interaction and social play. In infancy, strategies include positioning infants and arranging space; for toddlers, they include commenting positively on joint activities or encouraging ways to share space and materials.
- **Scaffolding Social Problem Solving.** Social Problems refer to the problems that arise among peers, defined as children under the age of 5. Most social problems among infants and toddlers involve objects (toys) and space. Infants are just learning that peers are people, rather than things. Caregivers may model "gentle touch" and put their hands over those of the child to help guide him or her to touch peers gently. Caregivers talk about peers as people with feelings. Caregivers may need to provide more physical support for social problem solving among infants (such as moving children or toys and redirecting children). As children get older, caregivers support children in solving problems with peers by offering them different strategies, such as finding a different but similar toy or taking turns. Caregivers talk about the perspective of others to help children understand that other children have feelings and these feelings may differ from the child's own.

C. Supporting language and literacy development

The parent-child literature provides a base for the aspects of caregiver behavior that support communication and language development, including both the verbal and nonverbal (gestural and preverbal) aspects of language. The constructs center around the quantity and quality of the language a caregiver directs toward the child, as well as the ways in which the caregiver invites child language and communication. The amount and diversity of parent language are related to child cognitive and language outcomes across a variety of ages (Dodici et al. 2003; Hart and Risley 1995; Huttenlocher et al. 1991; Huttenlocher et al. 2002). An increased number of utterances, variety of words, and types of talk (questions or sentences, for example) provide greater support for language development. Similar relationships are noted with Spanish-speaking children: the quantity and quality of mothers' talk prior to 18 months of age (both the number of utterances and the number of unique words) are related to 24-month-old child talk (Hurtado et al. 2008).

Diversity can be delineated further by the types of talk caregivers use with children, including questioning, the use of narratives, the caregivers' ability to tailor the talk to various developmental levels and purposes, and the use of many different words in context. For example, parents sometimes use questioning to initiate or extend conversation with children and may provide the answers for the infant or preverbal toddler. More frequent use of questions with infants and toddlers is associated with more positive language development (Tamis-LeMonda et al. 2001). Questioning also has been associated with memory abilities in 2- and 3-year-old children (Hudson 1990; Ratner 1984). However, since too much questioning may be intrusive for the child, it must be balanced with other forms of communication.

Narratives provide other forms of communication in a variety of features or types. For example, parallel language narrates or describes the child's activity during or immediately following the child's action. Children learn words and concepts more quickly when the adult discusses the function of the referent word (Booth 2009; Gelman 1990). Other types of

narratives—oral storytelling, accounts of what the child experienced in the past and other talk about things not physically present (decontextualized talk), or talk describing what will happen in the future (anticipatory talk)—also encourage children's language and memory skills (Heath 1986). Another aspect of language support involves conversational turn-taking (the back-and-forth interaction between caregiver and child, relying on responsive language). Turn-taking is facilitated when caregivers comment or ask a question, wait for a response, then follow up with a response that continues or extends the conversation.

The diversity of types of talk available in a caregiving environment may be important for child development. Tamis-LeMonda et al. (2001) examined the association between different dimensions of maternal responses and children's language development from 9 to 21 months. Maternal responses featured affirmations, imitations of child vocalization, descriptions (of an activity, event, or object), questions (about the activity, event, or object), play prompts or demonstrations, and exploratory prompts. The latter two categories could be in the form of a statement, question, or physical prompt. Maternal responses to children's play and vocalizations predicted the acquisition of language milestones, but different categories predicted different milestones at different ages. For example, use of affirmations and descriptions predicted early language milestones for children at 9 months, but not at 13 months. At 13 months, maternal use of vocal imitations and expansions were related to the achievement of language milestones. Also, while maternal use of questions at 13 months predicted the timing of children' first talk about the past, it did not predict any other language milestone within this sample. These findings suggest that in child care, where classrooms are expected to have children across a range of ages and abilities, the presence of different categories of caregiver responses should be considered.

Early language development is closely related to the development of future literacy skills. One avenue to support language is through sharing books, as they provide the opportunity to talk about things beyond the child's everyday experiences and learn to associate a word with a representation of an object or event, as well as with the object or event itself. Reading time with infants and toddlers presents opportunities for children to talk and share joint attention with an adult. Book sharing also provides an introduction to early social language, which further supports development; reading the words on the page is not enough to support children's language (Snow et al. 1998). Early book sharing also contributes to later literacy. Children who are read to by their parents early in life (younger than 6 months old) show greater interest in reading later (Lonigan 1994).

1. Dimensions of Caregiving Supporting Language Development

The following dimensions are important to assess when examining the quality of caregiver support for infant and toddler language development, and all are measured by the Q-CCIIT:

- **Use of Questions.** Caregivers can vary in whether they ask questions and how many questions they ask. A broad range of question types—yes-no, close-ended (naming and recall type) questions, open-ended questions (without a single answer), questions inviting more elaborate child responses, and questions that require thinking beyond recall—may better support development. In the use of questions, the wait time caregivers allow for a response is also important to capture. Young children take a longer time to process and organize a response.

- **Conversational Turn-Taking.** In maintaining the back and forth involved in conversations, caregivers may facilitate turn-taking that offers children multiple opportunities to communicate. In child care settings, caregivers often have many children in their care, some of whom are very verbal and sociable. It is important to ensure that all of the children, not just the extroverted children, have opportunities to engage in conversations.
- **Varied Vocabulary.** The caregiver's talk may be limited to common, nondescript words or pronouns. To support language development, varied vocabulary should include specific nouns, verbs, and descriptive words. At the highest level of quality, caregivers should provide some sophisticated words in context to support children's understanding.
- **Diversity of Talk.** Caregivers' use of narratives when interacting with infants and toddlers will differ greatly in frequency and type. The number of utterances and diversity of talk supports development, and capturing both is vital for examining support for different language milestones. Types of narratives include parallel language of children's activities, anticipatory talk about future events, positive comments or feedback (affirmations), descriptive talk about an object or activity, songs or poems, explanations (or cause-and-effect discussions) about the physical world or thoughts and feelings, reasoning, and decontextualized talk (that is, discussing people, places, or events that are not present).
- **Decontextualized Language.** Decontextualized talk is a more cognitively demanding type of language than description and questions about immediate objects, and it supports children's ability to represent ideas. We focus on its quality, not just its occurrence (as is done in the dimension above). For example, caregivers may use decontextualized language in limited ways ("Mommy is at work. She will be back later.") or more broadly ("Remember when we went on our walk and the butterfly landed on your finger?"). Caregivers may model it, scaffold with objects, or elicit it from children ("How did you get to school today?").
- **Extending Language.** Caregivers support language development by being responsive to children's communication (verbal or nonverbal) and then extending it by adding words or modeling full sentences.

2. Dimensions of Caregiving Supporting Early Literacy Development

Book Sharing. Experiences with books from birth to 36 months can influence language development, including children's exposure to more diverse vocabulary, more complex sentence structure, and different types of narratives. When reading to children, adults' language is often more complex and decontextualized than their language used in other activities.

Book sharing is also an important opportunity for joint attention and extending children's attention. The caregiver can read and discuss books in ways that fully engage children and extend their attention to the book, as well as lay the groundwork for a love of literacy.

When sharing books with infants and toddlers, caregivers also support language in the way they read the story (pointing to pictures, vocal tone, and asking questions); this can build meaning and understanding of print. The following list reflects dimensions of book sharing with infants and toddlers measured by the Q-CCIIT:

- **Engages Children in Books.** The caregiver uses a variety of ways to help children attend to, sustain interest in, and learn from books. The caregiver may use vocal tone, facial expressions and gestures (particularly to help children understand challenging vocabulary and complex texts), and encourage children to sit close and point to picture or turn pages. The caregiver may provide experiences related to the text; for example, giving children props or pictures to hold up at appropriate times or acting out the story or actions described in the text. Caregivers may ask questions, help children make connections to their own experiences, or explain what is happening in a story or text.
- **Uses a Variety of Words.** Books offer more opportunities to extend children's vocabulary than will be evident in most daily activities. The words used may be found in the text or added by the caretaker to give further description and should include use of nouns, verbs, adjectives, and adverbs. The nouns should extend beyond naming objects to naming feelings (for example, *frustrated, excited, afraid)*, parts of objects (for example, *steering wheel*, *limb* of a tree), or category words (*furniture, transportation, insects, fruit),* and should include the use of some challenging or sophisticated vocabulary (*nervously, ecstatic, valuable, hoarded, sputtered, swooshing, leapt, amazement).*
- **Uses a Variety of Types of Sentences**. The sentences used in sharing books are more complex than everyday language and may include language specific to the type of narratives (story, directions, poetry). Book sharing should include greater use of descriptive words (adjectives and adverbs), prepositions, and connective words (for example, *then*, *after*, *because*). For younger infants, longer and more complex sentences expose them to the rhythm of language. As children get older, caregivers should use pictures, gestures, connections to children's lives, and explanations in addition to the context of the story to help children understand more complex language.
- **Fosters a Positive Attitude Toward Books**. Caregivers can foster positive attitudes toward reading by reading frequently and inviting children to participate in book sharing in varied ways (for example, handle, look at, listen to, and talk about books) and make it a positive experience for all children. The types of books available (colorful picture books, familiar and novel stories, and informational texts) provide different experiences and opportunities to extend child interest and exposure to greater varieties of words and narratives.

D. Areas of concern in caregiving

Certain caregiving behaviors can have a negative effect on children's development. During an observation, negative behaviors often are highly salient and can help observers provide a more reliable overall measure of classroom quality if measured separately from positive behaviors. Although positive and negative factors often are strongly related to one another, and some researchers put negative behaviors on a continuum with positive, we follow other researchers' approach of investigating them separately (as in the Observational Record of the Caregiving Environment [ORCE] NICHD Early Child Care Research Network 2002b). Further, for the purposes of professional development and evaluating quality, it is important to measure both positive and negative behaviors.

The negative caregiving traits most consistently associated with adverse child outcomes include intrusiveness, negative regard, negative affect, harshness, over-controlling behaviors, and ignoring children or letting them wander aimlessly (Fuligni et al. 2004; Fuller et al. 2004; Ispa et

al. 2004). In addition, negative environmental effects, such as too much noise or overstimulation from television viewing (Als et al. 2004), and behaviors that threaten the health and safety of children (American Academy of Pediatrics et al. 2013) are also in this category. The Q-CCIIT examines whether there are any threats to children's emotional or physical safety.

The emotional health of young children is strongly related to the characteristics of the caregiving environment (Shonkoff and Phillips 2000; National Scientific Council on the Developing Child 2004; National Scientific Council on the Developing Child 2012). Caregiving behaviors that unduly stress a young child or that threaten positive emotional development of infants and toddlers have lifetime implications for children (National Scientific Council on the Developing Child 2004). When children experience child abuse, serious neglect, or exposure to violence, the stress response system activates. Frequent and strong activation of the stress response system can "result in the permanent disruption of brain circuits during the sensitive periods in which they are maturing." (National Scientific Council on the Developing Child 2007, p.8)

The American Academy of Pediatrics recommends no media use for children under the age of two and limited viewing for children over the age of two. Associations have been noted between the amount of time spent watching television and subsequent behavior problems, learning difficulties, sleep problems, and obesity (Brown 2011; Certain and Kahn 2002; Garrison and Christakis 2012). Despite popular beliefs that television provides exposure to richer vocabulary, television viewing among infants and toddlers is associated with poorer language and cognitive development (Duch et al. 2013; Tomopoulos et al. 2010). When toddlers watch television, the content of the viewing also matters (Garrison and Christakis 2012; Linebarger and Walker 2005). Even among preschoolers, viewing more violent media affects sleep (Garrison and Christakis 2012).

Careful supervision of children in safe and healthy environments helps to assure that children are physically safe. The American Academy of Pediatrics, American Public Health Association, and National Resource Center for Health and Safety in Child Care and Early Education (2013) collaborated in providing guidelines for child care settings in order to assure the physical safety and health of young children. The Q-CCIIT captures the presence of unsafe features in the environment, poor supervision of children, and absence of healthy practices, such as good sanitation and healthy nutrition.

When observing interactions with caregivers throughout the visit, information about the frequency and severity of negative caregiving behaviors is collected:

- **Overall Level of Chaos**. Disorganization of the room and activities, high noise levels, inconsistent and unclear limits, inconsistent caregiving, and frequently distressed children all contribute to high levels of chaos.
- **Physical Harshness.** Caregivers may interact in a rough or abrupt manner with children or use physical forms of discipline.
- **Verbal Harshness**. Caregivers' verbal interactions with children may be harsh in tone or volume of voice, or in what is said (yelling, harshly telling children to "shut up," verbally cutting children off, sarcasm, or calling children derogatory names).

- **Restricting Children**. Caregivers may restrain children for extended periods of time or intervene to stop a child from moving somewhere or doing an activity—physically, verbally, or removing objects—for reasons other than the child's safety.
- **Mismatch Between Caregiver Affect and Communication**. Caregivers' affect and actions may appear "off" (appearing fake) or not in sync with children's needs (for example, overly cheerful when the child is crying) or the communication (smiling when reprimanding a child). The mismatch in communication confuses children and can negatively affect the development of a trusting relationship.
- **Caregiver Singling Out Children**. Singling some children out for preferential positive treatment (in ways not related to individualization for needs) can communicate to other children that they are not as important or valued. Alternatively, some children may receive only negative attention while other children receive mainly positive attention.
- **Children Ignored**. Caregivers may ignore one or more children by not paying attention to or responding to children's verbal or nonverbal bids for attention, or by ignoring peer interactions or disputes that require adult intervention.
- **Children Unoccupied**. Caregivers may not attempt to engage children who appear bored or restless. Children may wander the room or lie or sit unoccupied for extended periods.
- **Children Overwhelmed**. The caregiver may interact in a way that is too intense for the children. For example, he or she may bombard children with questions, give children directions that have many steps or are unclear, play music too loudly, or provide many activities and toys at once. Evidence that children may be overwhelmed by the caregiver's behavior include their withdrawing from the activity or group, crying/fussing, or turning/moving away from the caregiver.
- **Children Stressed by the Demands of the Environment**. Examples of children exhibiting stress include frequent crying (without an apparent cause), fussing or whining, withdrawing to a quiet area, turning away from others, and frequent hitting or biting behaviors. The environmental factors that can prompt those behaviors are varied—noise, disorganization, intensity of interactions, or overly demanding activities.
- **Adult Television Use.** Adult television includes any television show or video not specifically designed for children, even when it is playing in the background.
- **Child Media Use**. The American Association of Pediatrics recommends that children under age 2 have no "screen time." If all children in the setting are over 2 years of age, the Q-CCIIT considers 30 minutes or less of educational media or television with the caregiver as appropriate use of media. Extended use or developmentally inappropriate media (for example, violent cartoons) would be considered an area of concern.
- **Poor Supervision of Safety**. The level of supervision caregivers provide ensures children's basic health and safety. Broadly, an observation tool should document when caregivers do not notice or intervene to stop children from engaging in unsafe activities, such as allowing them to play with dangerous tools or climb shelves, or leaving a child unattended when eating or near water (for example, a tub, pool, or pond).

- **Unsafe Environment.** Even when caregivers are supervising well, hazards may be present in the environment that pose a threat to young children, such as uncovered outlets, choking hazards, unanchored shelving, equipment in disrepair (for example, jagged edges or splinters on outdoor equipment), or open access to roads.
- **Sanitary/Healthy Practices Not Followed.** Healthy practices include providing appropriate clothing and sun/weather protection for outdoor activities, providing adequate healthy food and water, washing hands and surfaces, meeting diapering/toileting needs, and ensuring clean air (no secondhand smoke in the environment).

E. Additional information about the environment

Although the focus of the Q-CCIIT is on interactions and relationships, our conceptual framework notes that structural features and the environment can influence caregiving quality. Research indicates a relationship between global quality measures and measures of interaction (Forry et al. 2012; NICHD Early Child Care Research Network 2002a). In consultation with our expert panel, we included some aspects of the environment that may be important to include in examining caregiving interactions. Aspects of the environment addressing health and safety are captured in the areas of concern. We added items that capture professionally-recommended practices around the use of space and time such as , availability of a quiet area that children can access, the organization of the environment, , and the diversity of activities offered.

Collaboration with families is very important for supporting children's development (Ayoub, Vallotton, & Mastergeorge, 2011; Forry et al. 2011), particularly in these early years when children cannot communicate effectively. Although we knew that in some settings it would be challenging to observe, we piloted an item examining evidence of the support that caregivers provide to parents. We knew that a measure focused solely on family-provider relationships was also being developed in another project, so our efforts in this direction were not intended to be comprehensive.

Although these items about the environment and collaboration with parents are coded during the Q-CCIIT observation, they are complementary to the measure and not considered part of the scale.

- **Activity Schedule Balances Types of Activities.** Caregivers organize the day in such a way that allows for active times but also time for quiet activities and rest, indoor and outdoor activities (as permitted by weather and safety), and ample transition time. There is evidence of a schedule in the balance of activities across the day and through caregiver's actions (for example, following quiet inside activities with some time outside in vigorous play.)
- **Quiet Area.** A quiet area is an available and accessible space to which children can go when they need to relax, calm themselves, or otherwise need time alone. Non-mobile infants may be placed in a quiet area or a swing (for short periods) or given comfort objects to support self-regulation and avoid overstimulation. Examples of quiet areas include child-sized rockers, beanbag chairs, mats on the floor, or a collection of pillows in a separate space. The quiet area should be out of the main traffic pattern and apart from active play. In addition, it should have a pillow or something soft (stuffed animals, blankets, etc.) for the children to use to be comforted or relax.

- **Organized Caregiving Space.** The caregiving space should provide support for positive interactions, independent play, and self-regulation. Features of an organized environment include areas with adequate space for children to safely play together (caregivers may limit the number of children in a space), objects grouped together in activity areas so children know where to find materials, toys accessible for infants and toddlers, and a furniture/space arrangement that allows caregivers to have children within their visual line of sight.
- **Caregivers Are Supportive of Parents.** Parents are the child's primary caregivers, and interactions should recognize that relationship and treat parents as partners. Caregivers should make the setting a welcoming place for parents and greet parents (and children) when they arrive. Communication should be frequent and positive, and involve listening as well as sharing information. Caregivers may talk with parents individually upon arrival, or at the end of the day. They may exchange a notebook or activity sheets about the child's day, highlighting its positive aspects and celebrating growth and learning. When parents remain in the classroom, the caregivers support them by commenting positively and reinforcing their positive caregiving practices (for example, "She really cuddles close and calms down when you hold her like that"). Caregivers model good caregiving practices and may explain why they interact in certain ways (for example, "I need to give a one-minute warning to prepare the children for the transition to snack. It makes it easier for them to let go of the activity they're involved in.").
- **Varied Activities.** Throughout the day, a variety of activities should be available to engage children, including opportunities for different types of sensory and motor play, exploration of books and toys, and representational play.

F. Summary

In this chapter, we have described the dimensions assessed by the Q-CCIIT and the research base underlying them as important for supporting healthy development and learning. The Q-CCIIT includes measurement of positive support for social-emotional, language and literacy, and cognitive development, as well as noting caregiving behaviors that may impede development. The Q-CCIIT also includes some items that measure the quality of the environment in which these interactions occur.

This page has been left blank for double-sided copying.

III. Q-CCIIT INSTRUMENT AND ADMINISTRATION

A. Observation methods

The Q-CCIIT was designed to capture the dyadic nature of most interactions and represent the quality with which the caregiver manages multiple demands in the group setting. The Q-CCIIT observation tool includes 10-minute time samples that capture interactions occurring with a given caregiver and child/group of children at a given time, as well as global ratings based on the entire observation time. In each 10-minute sample (a cycle), the caregiver is rated based on the average experience provided to the children. When multiple caregivers are present in a setting, the observer focuses on a different caregiver in each cycle.

Although assignment of a primary caregiver for each child is recommended as best practice for infants, classrooms are structured in a variety of ways. A single caregiver may be available in a family child care setting; in a center-based setting, several caregivers may interact with each child throughout the course of the day. While caregivers within some classrooms display consistent ways of interacting with children, others may show diversity in caregiver style and quality of interactions. Thus, in settings with multiple caregivers, a focus on a single caregiver may not represent the average experiences of the children in that group. Alternatively, an observer cannot code the interactions of multiple caregivers and children simultaneously when they occur in different parts of the classroom. When the classroom is the focus of measurement, rather than an individual caregiver, observations should include all adults who provide direct care during the observation period, focusing on a single caregiver in each cycle and alternating caregivers across cycles.

The Q-CCIIT observation should be conducted at a time of day that allows a sufficient period of time to observe the range of interactions that children experience. Because routines and other events elicit different types of interactions, the observation should be structured to capture a representative sample of caregiver behavior during routine events (for example, feeding, free play) and ensure that less frequent events of interest, such as book reading, are coded when they occur. When using Q-CCIIT, observers should note the activity context for the interactions in each cycle.

Measurement of interactions occurring naturally in the environment requires longer periods of observation, particularly when observing multiple caregivers. Observing caregiving behaviors across several different activities elicits the most accurate range of behaviors (Munson and Odom 1996). However, observations with other quality measures longer than 2.5 hours have resulted in lower estimates of quality (Hofer 2010), so standardization of observation time is needed across programs. To capture multiple caregivers across a range of events in a variety of naturalistic situations, we recommend an observation lasting approximately 2 hours. When used within a classroom for professional development, the cycles could be distributed based on classroom schedules to optimize use of observer time. When used to compare programs, the observation time should be the same across programs to ensure comparability of measurement. The Q-CCIIT psychometric field test and validation study collected six 10-minute time samples, with up to 10 minutes of coding between cycles, across a 2-hour observation period.

The Q-CCIIT allows observers to code 12 dimensions within each cycle while coding other dimensions across the entire observation period (Table III.1). During each cycle, observers take notes that provide the evidence for both the cycle ratings and the dimensions rated across the visit at the end of the observation.

Table III.1. Dimensions coded within cycles and across the visit

Dimensions coded within cycle	Dimensions coded across the visit
Responding contingently to social cues	Responding contingently to distress[a]
Responding to emotional cues	Supervising and joining in play
Building a positive relationship	Responsive routines
Supporting peer interaction/play	Classroom limits/management
Supporting object exploration	Sense of belonging
Scaffolding problem solving	Support for peer social problem solving
Concept development	Extending pretend play
Use of varied vocabulary	Giving choices
Use of questions	Explicit teaching
Conversational turn-taking	Features of talk: varied narratives, sentence length, cognitive demand
Extending children's language use	Use of decontextualized talk
Book sharing: words, sentences, engagement	Developing positive attitudes toward books
	Areas of concern; Extreme areas of concern
	Environmental support: balanced activity, quiet area, organized space, partnering with parents, varied experiences

[a] Responding to distress was coded within a cycle during the field test, but distress was not observed very frequently and so was moved to Across the Visit.

To ensure more comparability across settings, the Q-CCIIT observation tool includes a semi-structured activity. The observer asks the caregiver to share the book *Good Night, Gorilla* (Rathman 1996) with one or more children.[5] This approach allows assessment of how caregivers invite children to participate in an adult-selected activity, how they transition to and structure the experience, and how they support development within this context, particularly the support for language and literacy development. *Good Night, Gorilla* has few words, and most of the story is told through the illustrations, offering a rich opportunity for individualized language support. Sharing books offers greater opportunities to support language than most play interactions and ensures that observers are able to see this opportunity in each environment (as it may occur at a different time of day and be missed).[6]

[5] At the start of the visit, caregivers are given the book *Good Night, Gorilla* and asked to share it with a child or children during the observation. No additional directions are provided. If asked, observers tell caregivers to share the book in whatever way they typically would share a book with children.

[6] In classrooms where book sharing does not occur regularly, the request to share a book with a child or children may inflate slightly the estimate of quality of the classroom in support of language and cognition beyond the typical quality of experiences in that classroom.

Using their notes to inform ratings, observers rate the majority of items on the Q-CCIIT on seven-point holistic rubrics that include four anchors describing how that dimension would look at different levels of quality. When deciding ratings, observers read descriptions, beginning with the first anchor, and select the description that best represents the average experience of the children with the caregiver during that cycle. When a caregiver's behavior differs across children or falls between two anchors (for example, caregiver provides more than the description included in a 5, but does not demonstrate all of the behaviors indicative of a 7, the observer assigns the intermediate rating, in the example, a 6).

Some items are not rated in every cycle. The Book Sharing items are rated only if the book sharing experience is at least two minutes long. The Support for Object Exploration and Support for Problem Solving are rated "not applicable" during interactions such as diapering, feeding, and book sharing.

The items in Areas of Concern are rated on a three-point frequency scale: (1) never occurs, (2) sometimes or briefly occurs, (3) frequently occurs or occurs for an extended duration. In addition, observers note if any incidence of the behavior was extreme, thus posing a more serious threat to children's well-being.

The Environment items are rated on a four-point scale indicating how characteristic the behaviors are in that setting. They were tested but not maintained as part of the final measure.

B. Administration

The Q-CCIIT rating form is the record of a given observation. It is organized to capture the following:

- Information about cycles (for example, start and end time, caregiver observed, interaction type, number of children awake during that cycle).
- Checklists of types of talk (to inform ratings) and concepts.
- Items rated each cycle.
- Items rated across the visit.
- Notes on the visit (for example, unusual events).

In the psychometric field test, we also collected overall ratings of each caregiver and the classroom, and a rating of how child-centered the classroom seemed to be, but these were used for research purposes to help us better understand how observers were using the rubrics and are not considered part of the Q-CCIIT observation tool.

The ratings collected during the cycles are based on the interactions occurring with a single caregiver during the 10-minute observation. The across-the-visit ratings are based on the entire observation period and the average experiences of children in that classroom across caregivers. When a classroom or home has more than one caregiver, we observe each caregiver[7] for at least

[7] The Q-CCIIT defines a caregiver as any adult who interacts with children in the role of supervisor or educator for at least one observation cycle (10 minutes).

one cycle to obtain an estimate of classroom quality. The Q-CCIIT observer follows each caregiver around the classroom to capture the average experience of the children in the care of that caregiver. We suggest that observers begin with the lead caregiver and rotate among the additional caregivers during subsequent observation cycles. Each caregiver should be observed during at least one cycle. In FCC homes with mixed age groups, at least one age-eligible child should be present during each cycle. The User's Guide (Atkins-Burnett et al. 2014 provides additional details on coding items, such as what to do when caregivers join or leave during a cycle, or when multiple caregivers are present.

IV. SCORING, INTERPRETATION, AND USE OF THE Q-CCIIT

In this chapter, we first describe how to score the Q-CCIIT. We then discuss issues in interpretation of Q-CCIIT ratings, including (1) accounting for variation both across and within caregivers and (2) determining whether to compare ratings to a criterion for performance or to the Q-CCIIT psychometric field test sample results. Finally, we discuss the potential use of the Q-CCIIT for three purposes—professional development, research, and evaluation—and the implications for interpretation of each use. To date, only psychometric work has been conducted with Q-CCIIT. This chapter focuses on functional use of the instrument, based on our psychometric analyses. For more detail about the results of the analyses, please see Chapter VII.

A. Scoring the Q-CCIIT

The Q-CCIIT provides 4 scale scores: Support for Social-Emotional Development, Support for Language and Literacy Development, Support for Cognitive Development, and Areas of Concern. Within the first three scales, observers rate some items in each 10-minute observation cycle, while other items are rated across the visit. When selecting a method for scoring the Q-CCIIT, we wanted to ensure that future users could calculate scores without special software. For items rated in each cycle, we first calculate the average rating across cycles and rounded it to the nearest whole number. These averages are then used as the score for those items, combined with the scores for the across-the-visit items from that scale, and we calculate the mean of all of the valid items in each scale (see Table IV.1). [8] For programs using the Q-CCIIT for professional development or self-monitoring, scores can be computed by hand using a scoring sheet. For researchers or evaluators working with larger samples, a spreadsheet with formulas embedded or a statistical package will allow for computing scores based on large numbers of classrooms.

B. Interpreting the Q-CCIIT

When interpreting Q-CCIIT ratings, we must account for two key considerations. First, we must consider sources of variation in scores both across and within caregivers. Second, we must consider whether to compare ratings to a performance criterion or a reference group (that is, the Q-CCIIT psychometric field test results). In this section, we discuss each of these considerations in turn. In the next section, we discuss the implications of these considerations for the different potential uses of the Q-CCIIT: professional development, evaluation, and research.

1. Sources of variation

During the Q-CCIIT field tests, scores varied across caregivers in a classroom and across activities and children for a single caregiver. Below, we discuss these sources of variation and their implications for interpretation.

[8] Some items may be missing if all children in the group are young, non-mobile infants (for example, Classroom Limits and Management, Extending Pretend Play) or if there is no distress during the observation period.

Table IV.1. Items by Q-CCIIT scale[9]

Support for social-emotional development	Support for cognitive development	Support for language and literacy development	Areas of concern[a]
Responding contingently to distress	Supporting object exploration	Use of varied vocabulary	Physically harsh
Responding contingently to social cues	Scaffolding problem solving	Use of questions	Verbally harsh
Responding to emotional cues	Concepts (diversity)	Conversational turn-taking	Restrict children
Building a positive relationship	Extending pretend playa	Extending children's language use	Affect mismatch
Supervises or joins in play and activities[a]	Explicit teaching[a]	Engaging children in books	Favoritism
Responsive routines[a]	Giving choices[a]	Variety of words (book sharing)	Ignore children
Classroom limits and management[a]	Supporting peer interaction/playa	Variety of types of sentences (book sharing)	Poor safety/supervision
Sense of belonging[a]	Supporting social problem solving[a]	Features of talk[a]	Overwhelm children
		Talk about things not present[a]	Children stressed
		Positive attitudes toward books[a]	Adult TV
			Child media
			Level of chaos
			Count of extreme behavior

[a] Rated across the visit.

[9] The scales were identified using exploratory and confirmatory factor analyses described in Chapter VII.

- **Variation Across Caregivers.** In the Q-CCIIT field test classrooms with more than one caregiver, we focused on different caregivers across cycles, observing each caregiver for at least one ten-minute cycle. Both during our development work and in the psychometric field test, our observers noted that caregivers in a single classroom sometimes varied greatly in the level of quality caregiving that they provided to children.[10] Obtaining a classroom-level estimate of quality required data collection across caregivers. Depending upon the purpose, users of Q-CCIIT will need to decide whether to focus on a single caregiver or multiple caregivers.
- **Variation Across Children for a Single Caregiver.** In our work on the Q-CCIIT development and field test, we noted that some caregivers did not provide the same quality of care and developmental support to all children. Observed differences in how some caregivers interacted with children of different temperaments and abilities was a topic of discussion in debriefing meetings. For example, it was challenging to rate a caregiver who interacted very little with or ignored one child who stayed with the caregiver throughout the cycle, when that same caregiver was very responsive to two children who stayed only briefly during the observation cycle. Children who were more outgoing and engaging might receive more attention and be involved in more conversations and activities than children who were more reserved, less mobile, or less verbal. In classrooms with a wider age span, younger infants were less likely to receive attention than toddlers. For some caregivers, the particular child (or children) involved in the interactions appeared to be a determining factor in the nature and quality of interactions. This may be especially true of caregivers working with toddlers, who can be more demanding.[11] For example, a caregiver may score positively for interactions with a child who is highly verbal and outgoing, since that child may make obvious positive social bids for caregiver attention. In contrast, the same caregiver may not notice nonverbal social bids made by a child who is less verbal or shy, thus scoring lower for interactions with that child. Although caregivers working with infants showed less variability and more consistency across interactions with different children, a particular infant may still be a determining factor in the nature and quality of interactions for some caregivers (such as a caregiver interacting with a healthy, happy infant versus an infant who is fussy from teething). This presents some challenges for observers when rating caregivers who are more variable in the quality of the interactions. In settings with multiple infants and/or toddlers, the observer needs to rate based on the average experiences of the children rather than on the experiences of a single child.
- **Variation Across Activities for a Single Caregiver.** In the Q-CCIIT field test, we also found that a single caregiver's ratings varied across different activities. Especially in the areas of Support for Cognitive Development and Support for Language and Literacy Development, the level of support for development that caregivers exhibited varied based on the content of the activity. For example, in Support for Language and Literacy Development, caregivers scored higher for book-related experiences, such as reading *Good*

[10] At the end of the visit, observers rated each caregiver from 1 to 7 in terms of overall quality of care provided. The variance within a classroom was as great as the variance between classrooms. The observed caregivers included "floaters" who might be in the room providing care for only 15 minutes while the primary caregivers took a break.

[11] Our stability estimates (test-retest) were weakest in toddler classrooms (n=14), particularly for items in Support for Social-Emotional Development and Support for Language and Literacy Development (see Tables D.1 and D.2 in Appendix D.

Night, Gorilla, than for other classroom activities.[12] Some of the items in Support for Cognitive Development cannot be rated during specified activities. For example, support for object exploration is not rated during book-sharing and some routine activities. Thus, depending on how the Q-CCIIT is being used, users may need to collect data across multiple activities to provide a reliable estimate of caregiver quality.

2. Interpretation: Performance Criteria vs. Reference Group

Ratings on the Q-CCIIT may be interpreted in relation to a specified performance criterion or comparisons to other classrooms. Below, we discuss the implications for interpreting Q-CCIIT scores when using different points of comparison.

- **Comparing Q-CCIIT Scores to a Criterion for Performance.** The Q-CCIIT may be used as a "criterion-referenced measure" where scores are interpreted in relation to a criterion for performance—in this case, the criterion encompasses what a caregiver should be doing to foster high-quality interactions with children, as represented by the behaviors described in the item descriptions and anchors 5 and 7 for each item. Caregiver performance may be assessed for strengths and weaknesses on each individual item in the measure as well as within and across domains. Caregivers or programs may select different practices and set professional development goals for improvement in that area. For example, a caregiver might set a goal of improving skill in extending children's language use. In addition to rating the caregiver on that particular item on Q-CCIIT, a coach or mentor might note during which types of activities and with which children the caregiver is more successful at extending language use. Other related items on the Q-CCIIT such as use of questions, use of varied vocabulary, and conversational turn-taking would also support understanding of quality in this area.
- **Comparing Q-CCIIT Scores to the Q-CCIIT Field Test Results.** The Q-CCIIT may be compared to results from other data samples—in this case, the Q-CCIIT field test results. Here, caregivers are compared not to a set of behaviors that they should be performing but to the average scores obtained in the field test sample. Although the field test results are based on a national sample[13] and we found a range of quality across the programs, the sample was a convenience sample not intended to be representative of early care and education programs in the United States. We did not randomly select centers and FCCs but looked for centers and FCCs within a geographic radius in each state, obtaining contact information from resource centers and local Early Head Start programs. Often the resource and referral center would provide information about our study to the programs directly, and we knew the names only of those who volunteered for the study. Thus, our sample results may not capture the full range and average quality found in caregiving in the United States. Nonetheless, the information from the national sample in the field test about caregiver quality on specific items with different age groups (Appendix C) can provide a point of reference for interpretation and may help in understanding caregiver performance or classroom quality. The results may also inform the selection of professional development targets by indicating which practices are easier and which are more challenging for caregivers to implement with quality, as well as the variation found in the frequency of implementation in relation to the

[12] The items associated with book-sharing had the lowest item difficulty. See Chapter VII for additional detail.

[13] Field test included observations of centers and FCCs in 15 states across 10 regions in the United States.

age level of the children in the classroom. The results described by setting (FCC and center; infant versus toddler classrooms) provide a point of reference for the naturally occurring differences in implementation of practices in different settings.

- **Timing of Observations.** The field test observations were conducted in the morning, usually after the primary caregiver had arrived. Scores on some of the Q-CCIIT items (such as items in Support for Cognitive Development) might be lower if classrooms are observed in the afternoon, when significant time may be devoted to napping and/or less experienced caregivers may be on duty.
- **Accounting for Differences in Items Per Scale.** Scale scores in some classrooms may be based on fewer items. Some items are not rated for very young non-mobile infants (for example, Classroom Limits and Management, Extending Pretend Play). Similarly, in classrooms where there is no child distress, caregivers may not have a rating on "Response to Distress." In interpreting scale scores, it will be important to consider the number of items used to estimate the score. In addition, in the areas of Support for Cognitive Development and Support for Language and Literacy Development, some items indicate different criteria for younger infants (less than 8 months of age). For example, conversational turn-taking does not have to be verbal with very young infants, but the caregiver must use words with children older than 8 months in order for the turn-taking to be considered a conversation.

C. Potential uses of the Q-CCIIT

Although to date only psychometric work has been done on the Q-CCIIT, the Q-CCIIT was designed with three primary goals for future use: professional development, evaluation, and research.[14] Below, we discuss each potential use of the Q-CCIIT and special considerations for each use.

1. Professional development

One potential use of the Q-CCIIT is to inform professional development efforts. To target professional development most effectively, we need to understand how individual caregivers support children across a range of activities as well as across the developmental spectrum and range of temperaments found in the classroom. Therefore, we recommend observing individual caregivers across a variety of activities and with all of the children in their care. In some cases, observations from different days may be helpful in looking at consistency or capturing interactions with different children. Specifically, we recommend observing a minimum of four cycles across different activities (see Appendix F for further information about the level of reliability for different numbers of cycles).

When used for professional development, Q-CCIIT scores typically would be compared to a criterion for performance, assessing caregiver performance on each item within the measure and identifying strengths and weaknesses both within and across domains. The patterns in performance may inform the selection of professional development targets or goals that are feasible for an individual caregiver. Caregivers (or their supervisors or mentors) might examine the cycle ratings and consider how consistent they are in providing support for each domain, as

[14] The Q-CCIIT has not yet been tested for the potential uses of professional development, evaluation, and answering research questions.

well as which activities challenge them in providing support. For example, caregivers may note that they provide very limited language during routine types of activities or that they do not take advantage of opportunities to support problem solving. Immediate professional development intervention efforts should be focused on any Areas of Concern noted, such as a caregiver demonstrating physical or verbal harshness, ignoring a child or children for extended periods of time, or threatening a child's basic health and safety (both physical and emotional health).

The Q-CCIIT field test results might also inform the professional development efforts. As mentioned above, although the field test sample was not nationally representative of centers and FCCs, the information from the field test about average caregiver quality on specific items with different age groups (Appendix C) can be used as a point of reference for interpretation and may help in understanding caregiver performance and selecting professional development targets. The field test results indicate which practices are easier and which are more difficult for caregivers to implement with quality, and how this implementation might vary based on ages of the children.

Using the Q-CCIIT for professional development entails interpreting scores on individual items. (By contrast, using the Q-CCIIT for evaluation or research typically involves interpreting scale scores.) Overall, ratings above 4 on an individual item reflect high quality caregiving practices that are responsive to infants and toddlers and provide support for their development. Based on data from the psychometric field test, we would expect that across classrooms, scores generally are highest on Support for Social-Emotional Development, followed by Support for Language and Literacy Development, with Support for Cognitive Development having the lowest scores. Overall, classrooms serving only infants score lower on average than classrooms serving toddlers or mixed age groups (FCCs).

2. Evaluation

The Q-CCIIT also may be used as an evaluation tool. For evaluation purposes, we suggest assessing the average experience of children across caregivers in a classroom rather than focusing on a single caregiver (that is, assessing across caregivers, children, days, and observers). The field test study administered the Q-CCIIT in this way when evaluating its psychometric properties. However, capturing the average experience of children in a classroom can be challenging. Specifically, with variation both across and within caregivers, rating a classroom will require a focus on different caregivers across cycles.

Within a single cycle, it is important to focus on a single caregiver at a time in order to limit the cognitive demand on the observer and provide more reliable estimates of quality. During the field test, it was difficult for observers in some classrooms to compute an average rating across how a single caregiver interacts with different children, even within the 10-minute cycle. For example, when looking at a particular caregiver, interactions with some of the children may resemble the behaviors described in a quality rating of 5, while interactions by the same caregiver may ignore another child or engage in behaviors with other children that are better described in a rating of 1. When rating this caregiver's interactions, the observer needs to consider the number of children, length of the interactions, and the frequency of interaction opportunities across 10 minutes.

The Areas of Concern play a critical role in interpreting the Q-CCIIT, often providing a signal of the variability in interaction across children[15] or indicating "red flags" that must be addressed for health and safety.

Given the complexity of rating caregiver behaviors across activities and children, evaluative uses of the Q-CCIIT would require checks over time for consistency on how raters weigh different behaviors in making their ratings.[16] In light of these challenges, we recommend collecting observations across different days (reflecting different content and different children) and different observers. This approach is consistent with findings regarding quality measures in elementary schools. The Measures of Effective Teaching (MET) project on research and evaluation of teaching quality recommended classroom observations conducted on four days by four different observers to obtain reliable classroom estimates (Kane and Staiger 2012).

The Q-CCIIT psychometric field test—with the limitations discussed above—can provide a point of reference for different areas of support. When comparing Q-CCIIT scores to the field test for evaluative purposes, observations should be conducted in the morning after the primary caregiver arrives. As mentioned above, scores on some of the Q-CCIIT items may be lower if classrooms are observed in the afternoon, when significant time may be devoted to napping and/or less experienced caregivers may be on duty.

3. Research

Finally, the Q-CCIIT may be used as a research tool. The research question will determine whether the Q-CCIIT should be interpreted across caregivers, activities, children, and/or observers as well as whether to compare Q-CCIIT results across samples or interpret results relative to a performance criterion.

We envision several different uses of the Q-CCIIT for research purposes. The Q-CCIIT could be used to explore the relationship between quality of caregiving and different curricula or approaches to providing caregiving. The Q-CCIIT allows for a detailed analysis of differences in caregiving approaches. For example, in the area of Support for Cognitive Development, we rate a caregiver on items for both play-based interactions, such as representational play and object exploration, and caregiver-directed interactions, such as explicit teaching (intentional teaching and planned experiences) and concept development.

Researchers might use the Q-CCIIT to examine the quality of different professional development initiatives as well as how Quality Rating and Improvement Systems (QRIS) influence classroom quality. In addition to an overall quality rating, the Q-CCIIT provides scores for scales that look at support for the different domains of development. Professional development efforts may benefit from information about which dimensions of caregiver interaction are improving. To date, QRIS efforts have only had environmental rating scales and structural features as quality metrics. The Q-CCIIT offers a measure of caregiver interaction as well as indicators of the safety of the environment (Areas of Concern).

[15] In particular, Areas of Concern items about singling out or ignoring children will provide information about the consistency of caregiving.

[16] This complexity also makes a checklist an inappropriate tool for evaluation, as a caregiver may be having many good interactions, but only with some children.

Finally, researchers could extend the psychometric work already conducted with the Q-CCIIT. For example, researchers might replicate the findings using diverse and nationally representative samples, or provide further evidence of validity. The Q-CCIIT field test provided evidence of content and construct validity, and convergent/discriminant validity with other measures of quality. Additional work is needed to look at Q-CCIIT's relationship with child outcomes—in particular, the association of Q-CCIIT with changes in child outcomes across time.

D. Summary

Like any other classroom observation measure, the Q-CCIIT faces the challenge of accounting for variation across caregivers in a classroom and across activities and children for individual caregivers. To date, only psychometric work has been done on the Q-CCIIT; however, as detailed in Chapter VII, the Q-CCIIT's strong reliability, sensitivity to variation in caregiving, and evidence of validity support its ability to provide estimates of quality across and within caregivers and suggest its utility for the potential uses of professional development, evaluation, and research. The caregiving behaviors vary in difficulty and are based on empirical evidence of caregiving behaviors related to child outcomes. The Q-CCIIT offers the opportunity to identify strengths and challenges in caregiving in a variety of settings and the potential to test different approaches for improving caregiving for children.

V. DEVELOPMENT PROCESS

We used a four-phase approach to develop, operationalize, and refine the Quality of Caregiver-Child Interactions for Infants and Toddlers (Q-CCIIT) measure and collect data on its psychometric properties. In this chapter, we describe the initial phase, comprising a literature review and the development of a measurement framework, followed by three data collection phases we refer to as the pretest, pilot test, and psychometric field test. The number of observations, geographic locations, and observers increased with each phase of data collection. With each phase, we refined the measure until we ultimately evaluated the psychometric properties of the final measure. We describe the literature review, pretest, and pilot in this chapter, the psychometric field test methodology in Chapter VI, and the field test findings in Chapter VII.

A. Phase 1: literature review and measurement framework[17]

The first phase of the Q-CCIIT project involved a literature review to provide a summary of the extant measures of quality appropriate for use in nonparental care environments that serve infants and toddlers, and to evaluate the degree to which these measures provided adequate measurement of important features of quality. The Q-CCIIT literature review identified existing measures of adult-child interactions in infancy and toddlerhood—both for parenting and child care settings. Nearly half of the reviewed studies used an author-developed observational measure or coding scheme rather than a published, validated measure. Measures of caregiver-child interactions tended to be developed to capture dyadic parent-child interactions, whereas measures of child care setting quality tended to be developed to capture overall quality. The most prevalent constructs covered by caregiver-child interaction measures included sensitivity/responsiveness, support for language and cognitive development, positive regard, positive affect, and negative regard. The least prevalent constructs covered included reciprocity, joint attention, detachment, and negative affect. The measures showed a range of strengths of association with children's cognitive, language, and social-emotional competencies. The literature review confirmed the interaction constructs that should be represented within the Q-CCIIT measure; the team also noted that indicators of these components may be operationalized differently based on the age of the child or variations in cultural backgrounds.

The Q-CCIIT team developed a measurement framework for the items and observation techniques used in the Q-CCIIT measure. In Chapter II, we described our approach to selecting key constructs that guided the measurement development. The team developed and refined items for the preliminary Q-CCIIT measure in early 2011 based on the literature review, measurement framework, conceptual model, discussions with experts in our Technical Work Group (TWG), and through review of video footage from early child care settings. We tested and refined the initial measure during an iterative pretest data collection.

[17] The citation for the literature review is as follows: Halle, T., R. Anderson, A. Blasberg, A. Chrisler, and S. Simkin. "Quality of Caregiver-Child Interactions for Infants and Toddlers (Q-CCIIT): A Review of the Literature." OPRE 2011-25. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2011. We provided the Technical Work Group and the Administration for Children and Families (ACF) with the measurement framework in October 2011.

B. Phase 2: pretest

The primary goal of the pretest was for trained Q-CCIIT observers to work iteratively with the measurement development team to test and refine the instrument and observation procedures. As depicted in Table V.1, we observed 20 center-based infant and toddler classrooms and 8 family child care homes (FCCs) with infants and/or toddlers in New Jersey, Maryland, Virginia, and Washington, DC. We carried out observations during the mornings to take advantage of opportunities to observe caregiver-child interactions under a variety of circumstances (for example, arrival, feeding, diapering). We categorized center-based classrooms as "infant classrooms" when 80 percent of the classroom comprised children between birth and 14 months of age, and "toddler classrooms" when 80 percent of children in the classroom were between 15 and 36 months.[18] Most of the FCCs we visited included children younger and older than 3 years of age (referred to as "mixed age" settings). Four of the classrooms were bilingual, with Spanish and English spoken. Observers included members of the measurement development team and professional staff with a range of backgrounds (bilingual, different educational backgrounds). Pairs of observers conducted the pretest observations to facilitate discussion of items among observers and video record observations. The measurement development team used the video recordings to apply and refine items when further developing the Q-CCIIT measure. The videos also became part of future training exercises. Observers participated in weekly debriefing calls with the measurement development team.

In the pretest, we found that interactions differed in infant versus toddler care settings. Typically, toddlers were more active and elicited more attention from the caregivers, while infants principally received primary care with more limited support for language and cognition. In mixed-age groups (including in FCCs), we noted that infants and toddlers could have very different experiences in the same setting, with the older and more verbal children eliciting more positive caregiving practices than the younger children. Estimates for almost all items were slightly higher in mixed-age settings, the majority of which were FCCs. The exceptions were support for object exploration (higher in infant classrooms) and scaffolding problem solving (higher in toddler classrooms).

Items within the measure were refined on a regular basis during the pretest in response to discussions with the observers and review of video recordings. After the pretest, experts from our TWG reviewed the items; they supported the selected constructs and representation of the constructs in the items, and suggested some refinements and the addition of items about the environment.

[18] For the psychometric field test, we changed our definition: children less than 18 months were categorized as infants and children 18 months or older as toddlers. This change was reflective of differences in age groupings used by state licensing rules for designating required child:caregiver ratio.

Table V.1. Participating classrooms, by phase

	Pretest (spring 2011)	Pilot test (winter 2012)
Total Q-CCIIT Observations	**28**	**60**
Center Classrooms	20	30
Family Child Care	8	30

C. Phase 3: pilot test

The goals of the pilot test were to test the almost-final instrument in the field and examine the initial psychometric evidence. During the pilot test, 12 observers (9 of whom had participated in the pretest) conducted the observations.[19] As depicted in Table V.1, from winter to spring 2012, these trained observers conducted 60 classroom observations, equally divided between FCCs and center-based classrooms. The observations occurred in New Jersey, Pennsylvania, Maryland, Virginia, and Washington, DC. Classrooms with at least two children between birth and 36 months were purposively selected for observations to ensure a variety of child care settings (center based and FCC); age range variety (infant, toddler, and mixed-age groups); and the cultural, linguistic, and economic diversity of the populations served. We categorized center-based classrooms as infant or toddler using the same criteria used in the pretest. A majority of the FCCs observed (23 out of 30) had classrooms comprising children with mixed ages. There were 14 infant classrooms, 19 toddler classrooms, and 27 mixed-age classrooms observed across settings. Fourteen classrooms were Spanish and English bilingual, one classroom was bilingual—English and another language—and in the remainder of the classrooms, the caregivers spoke only English. From the 60 classrooms, 107 caregivers completed a self-administered questionnaire. Ninety-five percent of these caregivers were women. Thirty-five percent were African-American non-Hispanic, 26 percent were white non-Hispanic, 23 percent were Hispanic, and 16 percent were "other" or did not report on race or ethnicity. The pilot data collection informed the updating of the Q-CCIIT items, training, and observations procedures for the final Q-CCIIT measure, which then was examined as part of the psychometric field test. Next, we present a brief summary of results from the pilot data collection.

D. Findings from the pilot test

The pilot data provided an opportunity to examine the draft Q-CCIIT measure's characteristics by subgroups of setting type (child care centers versus FCC settings) and age. The team conducted the following analyses using the pilot test data: (1) assessing inter-rater reliability of the measure, (2) testing different scoring approaches, (3) examining descriptive statistics at the item and scale levels for the full sample and by subgroup, (4) assessing internal consistency reliability and item-total correlations, and (5) assessing reliability and construct validity through item response theory (IRT) analysis.

[19] Observers were trained on the revised measure for three days and certified at the end of training based on agreement of their ratings with those of the measurement development team, using video from four different classrooms.

Inter-Rater Reliability. We conducted nine paired observations over a 12-week period. We examined adjacent agreement, averaged across the cycles, for four proposed scales: (1) Support for Social-Emotional Development, (2) Support for Cognitive Development, (3) Support for Language and Literacy Development, and (4) Areas of Concern. All scales had inter-rater agreement (within one point) of 90 percent or higher. The team examined checklists for exact matches, with percentage agreement averaged across cycles. The level of agreement reached 91 percent for Concept Development and 84 percent for Types of Talk, the two sets of checklists. Across the entire rating form (all ratings and checklists), the level of agreement was 92 percent.

Scoring Approach. In the pilot study, we tested different approaches to scoring and conducted a preliminary examination of proposed dimensions and scales. For items collected through the six observation cycles, we calculated the mean across all valid cycles and then created a mean score (including items rated at the end of the observation period) for each of the scales. For items that involved a higher use of "not applicable" (NA) ratings, we calculated the mean rating of the highest four cycles and found they were strongly correlated with the mean across all cycles (with correlations around 0.99). For Concept Development, we calculated the number of different concepts presented in at least one of the observation cycles and the frequency of discussion of concepts. We also checked for differences in scores for different classroom events (for example, feeding and free play).

Descriptive Statistics. We examined item-level descriptive statistics for the full sample and by setting type (center-based classroom versus FCC) and age (infant, toddler, and mixed). Centers had higher scores than FCCs on Support for Social-Emotional Development and Support for Language and Literacy Development but lower scores on Support for Cognitive Development. The scores on all scales were lower in infant classrooms than in toddler and mixed-age classrooms. We also noted the frequency of use of different items in the checklists Types of Talk and Concept Development.

Internal Consistency Reliability and Item Total Correlations. We computed the coefficient alpha for each proposed scale: Support for Social-Emotional Development (0.87), Support for Cognitive Development (0.84), and Support for Language and Literacy Development (0.92). With limited variance, the reliability estimate was lower but still acceptable for Areas of Concern (0.71).

We also examined the variation in the use of rating categories in an item and the item-total correlations for each item. Depending on the level of measurement, the item-total correlation was an r-biserial (for ratings and rubrics) or a point biserial (for dichotomous checklist items). In a few cases, we made revisions to the instrument based on the data. For example, in the Areas of Concern scale, we found that category 3 was very seldom used; therefore, we changed the rating scale to a three-point scale and added a separate box to capture extreme severity.

Item Response Theory (IRT) Analysis. We conducted preliminary IRT analysis for the pilot data of three of the proposed scales (Table V.2: Support for Social-Emotional Development, Support for Cognitive Development, and Support for Language and Literacy Development). With 60 classrooms, our analysis of the pilot data was only exploratory, with the goal of identifying problematic items and categories. Preliminary results showed that the ordering of item difficulties was consistent with theoretical assumptions. The item fit statistics were in the acceptable range. The reliability estimates for the scales ranged from 0.85 to 0.91.

The mean raw score and IRT score for a specific scale were highly correlated: the correlation was 0.98 for Support for Social-Emotional Development, 0.92 for Support for Language and Literacy Development, and 0.99 for Support for Cognitive Development. Inter-factor correlations among these scales ranged from 0.73 to 0.81 for the mean scores and from 0.75 to 0.80 for the IRT scores.

Overall, the pilot data collection provided further information for refining the items and observation procedures. In the next chapter, we describe the methodology for the final data collection conducted to document the psychometric properties of the Q-CCIIT measure, referred to as the psychometric field test.

Table V.2. Latent factors/scales and observed indicators proposed in analysis plan

Support for social emotional development	Support for cognitive development	Support for language and literacy development	Areas of concern[a]
Responding contingently to distress	Supporting object exploration	Types of talk	Physically harsh
Responding contingently to social cues	Scaffolding problem solving	Use of varied vocabulary	Verbally harsh
Responding to emotional cues	Extending pretend playa	Use of questions	Restricts children
Building a positive relationship	Explicit teaching[a]	Conversational turn-taking	Affect mismatch
Supporting peer interaction/play	Giving choices[a]	Extending children's language use	Favoritism
Support for social problem solving a	Supervises or joins in play and activities a	Engaging children in books	Ignoring children
Responsive routines[a]		Variety of words (book sharing)	Poor safety/supervision
Classroom limits and management a		Variety of types of sentences (book sharing)	Overwhelms children
Sense of belonging[a]		Features of talk[a]	Children stressed
		Talk about things not present[a]	Adult TV
		Positive attitudes toward books[a]	Child media
			Level of chaos
			Count of extreme behavior

[a] Items rated at the end of the observation, including all items in Areas of Concern.

This page has been left blank for double-sided copying.

VI. PSYCHOMETRIC FIELD TEST METHODS

The Q-CCIIT psychometric field test took place in 10 geographical clusters spanning 14 states and the District of Columbia from September through November of 2012.[20] We sampled purposively to obtain variation in care in each region and meet our targets for infant classrooms, toddler classrooms, and family child care homes (FCCs). The following additional criteria were used to inform our choice of sites:

- Region of the country
- State teacher-child ratios
- Racial and ethnic diversity
- Linguistic diversity
- Household income diversity

Given the challenges of recruiting FCCs that are not as familiar with classroom observations and research, we sought to maximize access to FCCs when choosing sites for the psychometric field test. We began by looking at Head Start Program Information Report (PIR) data from Early Head Start (EHS) programs associated with FCCs. We also gave priority to sites where we had a local contact who might assist us in gaining access to FCCs. In all cases, we sought to maximize diversity (and variability) in the sample.

The final sample in the field test included 400 classrooms[21] (110 FCCs and 290 center-based classrooms). The goal of the psychometric field test was to document psychometric evidence for the final Q-CCIIT instrument. As depicted in Table VI.1, during this phase, data collection activities were designed to support the analysis of the new measure's reliability and validity. This phase of the study included the following five activities:

1. Conducting Q-CCIIT observations in 400 classrooms.

2. Collecting information about the caregiver and classroom from caregivers in the 400 classrooms through the use of a self-administered questionnaire (SAQ) and classroom roster questionnaire.

3. Conducting test-retest reliability observations in 62 classrooms (32 center-based classrooms and 30 FCCs).

4. Conducting inter-rater reliability observations in 52 classrooms (41 center-based classrooms and 11 FCCs).

5. Conducting validation observations with three additional measures of classroom quality (Table VI.1):

[20] A separate memo summarizing the site selection approach was submitted to ACF in April 2012.

[21] A total of 403 classrooms were observed; however, three of the classrooms had fewer than five valid observation cycles (that is, observation cycles that lasted five minutes or longer) and thus were not included in the analyses.

a. Observational Record of the Caregiving Environment (ORCE) (NICHD Early Child Care Research Network 1996, 2002) in 119 classrooms (including centers and FCCs) and Infant/Toddler Environment Rating Scale-Revised (ITERS-R) (Harms et al. 2006) in 65 center classrooms.

b. Family Child Care Environment Rating Scale-Revised (FCCERS-R) (Harms et al. 2007) in 49 FCCs.

Table VI.1. Q-CCIIT psychometric field test observations

	Total	FCC	Infant	Toddler
Q-CCIIT observations	400	110	136	154
Test-retest	62	30	18	14
Reliability pairing	52	11	17	24
ORCE	119	41	36	42
ITERS	65	N/A	30	35
FCCERS	49	49	N/A	N/A

A. Reliability and validity methodology

Conducting the Q-CCIIT Observations. A total of 25 trainees[22] participated in a six-day, in-person training. Prior to training, trainees were asked to review written materials and complete a written quiz to assess their understanding of the material. The training included formal presentations, interactive activities, group video coding, and a live observation with a gold standard observer,[23] culminating in certification based on coding video segments from two classrooms. We certified trainees if they achieved at least 80 percent agreement (within one point) for items across the entire measure and further demonstrated at least 75 percent agreement for core subscales. Eleven trainees were certified using these criteria. Another 5 trainees met the overall criterion, but did not meet a subscale criterion. These 5 were considered "provisionally certified." All certified and provisionally certified field staff received personalized written feedback based on patterns we observed in their certification scores that indicated strengths and weaknesses. Provisionally certified field staff also had the opportunity to practice by scoring two video segments from two additional classrooms, comparing their scores against gold standard scores, and debriefing with gold standard observers by phone. Finally, 9 trainees did not meet the criteria at the conclusion of the training and did not participate in the psychometric field test.

Gold standard observers started to pair with individual field staff for in-person visits during the third week in the field. During these pairings, all field staff met criteria with the exception of one provisionally certified observer who did not remain on the project and never collected data

[22] Trainees ranged in age from 30 to 72 years old. Twelve of the trainees were Spanish and English bilingual. Forty-eight percent of the observers were Hispanic, 32 percent were African American, and 20 percent were white Non-Hispanic. All had bachelor's or graduate degrees. Most had conducted early childhood classroom observations previously (using measures such as the ITERS, the Early Childhood Environment Rating Scale (ECERS), or the Classroom Assessment Scoring System (CLASS).

[23] Gold standard observers were reliable observers during the pretest and pilot data collections described in Chapter V.

independently. Among the 16 trainees who were certified at or within a week of training, the inter-rater reliability averaged 79 percent agreement across all items at the end of training (ranging from 60 to 98 percent).[24]

Assessing Test-Retest Reliability. To evaluate the temporal stability of the Q-CCIIT, the same observer conducted a second Q-CCIIT observation within two weeks of the first observation in 62 of the 400 classrooms (32 center-based classrooms and 30 FCCs). We examined the Pearson correlations between the ratings across days, both overall and by scales. In addition, we examined the variability at the item level. The composition of the classroom—both the children present and the caregivers in the classroom—changed from one observation to the next in some of the observation settings. For example, some children did not attend the setting full time and so might be present for only one day, or substitute caregivers sometimes were in the room.

In addition to the day-to-day variability, we also used generalizability theory to explore the stability of the scores across cycles (generalizability or G-study) and estimate the number of cycles that need to be collected to obtain a reliable estimate of the caregiving quality (decision or D-study). The analyses involve a decomposition of the variance between classroom, cycle, item, and observer. Our design had several limitations, and the results are exploratory. The design was not balanced (for example, data included items that were missing validly, such as Response to Distress, across multiple cycles) and did not include multiple raters of the same classroom; also, classrooms were nested in observers. Given the limitations, we conducted the analyses using different methods in order to confirm our results: (1) with all of the data, (2) without items that had high levels of missing data, (3) with data imputed, and (4) using the Brennan software developed specifically to account for the unbalanced design.

Assessing Inter-Rater Reliability. We assessed inter-rater reliability in the field test in two ways. We scheduled paired observations between field staff in different geographic locations, as well as at least one paired observation carried out by each of the 15 field staff with a gold standard observer at weeks three or five of the field period. Using a seven- to eight-person observer team in each geographic area, we conducted 35 paired field staff observations combined with 17 gold standard paired observations (one or two for each field staff), resulting in a total of 52 paired observations.

We examined inter-rater reliability in the field by estimating adjacent rater agreement using an approach similar to certification reliability.[25] We estimated the adjacent agreement (percentage of items with agreement by subscale and overall), correlations between raters, and weighted Kappas. The weighted Kappas adjust the agreement for the possibility that one observer may have matched the other observer by chance. Weighted Kappas are affected by the prevalence of ratings, with more skewed rare behaviors demonstrating lower Kappas that do not reflect the level of agreement. We also used the G-study results to examine the amount of variance attributable to the observer when accounting for item and temporal reliability.

[24] At the end of training, some observers were certified provisionally and conducted practice observations in the weeks following training. They demonstrated reliability and were certified prior to conducting independent observations.

[25] For more details, see the certification plan delivered to ACF in July 2012.

Evaluating Content and Face Validity. The items for the Q-CCIIT were developed based on a review of the research literature on caregiving practices associated with more positive outcomes for infants and toddlers. The literature base included research on parent-child interaction; social-emotional, language, and cognitive development of infants and toddlers; and recommended practices in classroom management and group care. We also reviewed measures of the environment and literature about supportive environments to select environmental factors that should be evaluated when examining the quality of group care for infants and toddlers.

Our items were reviewed by a panel of experts in child development, child care, classroom observation, and home environment quality measure development. They confirmed the face validity of the items: that is, the items assess the practices and types of interaction that are supportive of infant and toddler development. In addition, the panel recommended the addition of the items about the environment to capture some of the context for caregiving (schedule, quiet space, organization, supportive interactions with parents, activities offered, and additional health and safety items for Areas of Concern).

Evaluating Construct Validity. We examined the construct validity of the Q-CCIIT using both confirmatory factor analyses (CFA) and Item Response Theory (IRT). We examined the fit of the data to the CFA model, both for the overall sample and by subgroups. We looked at the strength of the factor loadings and examined alternative models to identify and confirm the best model. Using IRT, we examined the reliability of the model and the hierarchy of items (map of item difficulties) in relation to theoretical assumptions about the construct.

Assessing Convergent and Discriminant Validity. One method for examining the validity of the Q-CCIIT measure is to examine its association with other classroom quality measures. We examined convergent and discriminant validity during the field period by having a second validation observer conduct an observation with a different observational measure in 233 of the 400 classrooms (90 FCCs and 143 center-based classrooms).

In selecting the validation measures, we considered the psychometric properties of the measures, including face validity, predictive validity (including demonstrated relations to child outcomes in research independent of the developer), and reliability.[26] We also considered the breadth of constructs covered in the measure and its relation to the constructs included in the Q-CCIIT measure, the degree to which the measure was well known and used by others in the field, its use with diverse populations, and its ease of administration. As a result of the Q-CCIIT literature review (Halle et al. 2011), and in consultation with the TWG and project officer, the ORCE and the Environment Rating Scales (ERS) (ITERS-R/FCCERS-R) were selected for use as the validation measures. The ORCE measures caregiver-child interactions, with a few items addressing language and cognition. The ITERS-R is designed for assessing the caregiver-child interactions and classroom environment in center-based settings, and the parallel FCCERS-R is for FCC homes. The validation measures each were administered by independent observers at the same time the Q-CCIIT was administered; however, the observation time was one hour longer for the ITERS-R and FCCERS-R validation measures (three hours each) compared to the Q-CCIIT (two hours). The following section provides more detailed information on each validation measure selected.

[26] A separate memo summarizing validation measure selection criteria was submitted to ACF in September 2011.

B. Validation measure: the ORCE

The ORCE was developed to assess caregiver-child interaction quality in early child care settings (NICHD Early Child Care Research Network 1996). Like the Q-CCIIT measure, the ORCE was designed for use in center-based and FCC settings, and thus provides a snapshot of the quality of care that infants and toddlers experience in different types of settings. Specifically, the ORCE measures caregiver behaviors identified in the research literature as contributing to children's cognitive, language, and social-emotional development. The focus is on interactions with a particular child (and in some cases, with other children in the classroom if the focal child is a member of the group) during a specific time period, rather than the average experiences of the children in the classroom. To capture developmental differences between children of different ages, separate forms are available for children who are approximately 6, 15, 24, and 36 months old. Although the age forms vary slightly to be age appropriate, they are designed to assess similar dimensions of quality; thus, there is considerable overlap in the caregiver behaviors observed at each age.

ORCE Observation Procedure. The ORCE observers completed two 44-minute cycles of the ORCE for each classroom. Each cycle consisted of three 10-minute segments, during which the observers alternated between 30-second observe and 30-second record intervals to assess the incidence of specific behaviors. Within a given 44-minute cycle, each 10-minute segment focused on a different target child. During the 30-second observe intervals, observers focused on the caregiver's behavior with the target child (and in some cases, with other children in the classroom); during the record intervals, observers completed the behavior checklist. Following each 10-minute segment, for 2 minutes observers completed qualitative ratings of caregiver behavior, which capture the quality and nuances of caregiver behaviors in relation to the target child's behaviors. During the final 8-minute period, observers took descriptive notes devoted exclusively to the qualitative ratings across all target children. At the end of this period, observers completed a final set of qualitative ratings, this time reflecting the collective experience of the target children as a whole. After observers completed two 44-minute ORCE observation cycles, they completed the end-of-visit ratings based on their overall impressions during the observation.

Expected Relationship of the ORCE and Q-CCIIT Measures. The ORCE originally was used to follow a single child in a caregiving setting and obtain an estimate of that child's caregiving experiences. Because the Q-CCIIT measures the average experience of the children in a given environment and our intent was to compare classroom estimates, we discussed different approaches to conducting the observation with the ORCE developers from the NICHD Study of Early Child Care (SECC). Based on their recommendations, we adapted the ORCE administration to follow three randomly selected children, rather than a single child in each setting. Although the ORCE is focused on the interactions of the teacher with individual children, and the Q-CCIIT includes both dyadic and group interactions, we expected relatively strong associations between the ORCE emotional-sensitivity scales and the Q-CCIIT Support for Social-Emotional Development (positive association) scale. The ORCE scale has only a few items addressing language and cognition, so we expected weaker correlations between the ORCE and the Q-CCIIT Support for Cognition and Support for Language and Literacy Development. We conducted validation observations with the ORCE in 119 classrooms (78 center-based and 41 FCCs).

ORCE Scales. The ORCE examines the characteristics of caregiving in two different ways:

1. The Qualitative Rating Scales capture the quality of caregiver behaviors (for example, the extent to which the caregiver tries to foster the child's learning).
2. The Behavior Checklist provides a record of the occurrence (or quantity) of specific caregiver behaviors (for example,, the frequency with which the caregiver responds to the child's vocalizations).

Qualitative Rating Scales. We conducted the qualitative ratings after each observation segment and then after each of three segments (one cycle) for up to three caregivers. Following the NICHD SECC scoring approach, we created the quality of care variables along the following nine dimensions by taking the average rating across caregivers both within segments and cycles, and then averaging across segments and cycles:

- Sensitivity/Responsiveness to Distress[27]
- Sensitivity/Responsiveness to Non-Distress
- Lack of Intrusiveness
- Lack of Detachment/Disengagement
- Stimulation of Cognitive Development
- Positive Regard for the Child
- Lack of Negative Regard for the Child
- Lack of Flatness of Affect
- Fostering Exploration[28]

Each dimension is rated along a four-point scale, with higher scores indicating higher quality: Not at all characteristic; Minimally characteristic; Moderately characteristic; and Highly characteristic.

In addition to these nine dimensions, three composite variables were constructed for the Qualitative Ratings. The overall qualitative rating is a composite defined as the mean of Sensitivity/Responsivity to Non-Distress, Lack of Detachment/Disengagement, Stimulation of Cognitive Development, Positive Regard for Child, Lack of Flatness of Affect, and Fostering Exploration. In the NICHD SECC, the Cronbach alphas ranged from 0.83 to 0.89 for the overall qualitative rating (NICHD Early Child Care Research Network 2000). Overall positive rating is a composite defined as the mean of Sensitivity/Responsivity to Non-Distress, Stimulation of Cognitive Development, Positive Regard for Child, and Fostering Exploration. Overall Lack of Negative rating is a composite defined as the mean of Lack of Detachment/Disengagement, Intrusiveness, Negative Regard for the Child, and Flatness of Affect.

[27] Available for the 6- and 15-month forms.

[28] Available for the 36-month form only.

Behavior Checklist. We constructed three composite measures from the Behavior Checklist. *Language stimulation* is a composite variable based on the mean occurrence across segments on the number of 30-second observation intervals in which the caregiver responded to the child's vocalizations or talk, read aloud to the child, asked the child questions, directed other talk to the child, and stimulated the child's cognitive development (for 6- and 15-month forms) or taught the child academic skills (for 24- and 36-month forms). In the NICHD SECC, the Cronbach alphas ranged from 0.88 to 0.92 for this measure. *Positive behavior toward child* is a composite variable based on the mean occurrence across segments on the number of 30-second observation intervals in which the caregiver shared positive affect (6- and 15-month forms), made positive physical contact, spoke positively to child (15- to 36-month forms), and engaged in mutual exchange (24 and 36 months). *Negative behavior toward child* is a composite variable based on the mean occurrence across segments on the number of 30-second observation intervals in which caregiver spoke negatively to the child, used negative physical actions, restricted child activities, or restricted the child in a physical container (6- and 15-month forms). Scores for the Behavior Checklist composites can range from 0 to 10.

Descriptives of the ORCE Scales in the Q-CCIIT Field Test. Table VI.2 presents descriptive statistics for the ORCE scales and the composites for the full ORCE observation sample in the Q-CCIIT field test. Appendix B presents descriptive statistics on the ORCE by subgroups. Classrooms were rated high on most of the qualitative rating scales, with means greater than 3.0 on the 4-point scale. The two exceptions were Stimulation for Cognitive Development and Fostering Exploration, for which the means were 2.12 and 2.39, respectively. Table VI.3 presents the information by program type and child age. The Overall Qualitative scores were higher for infants and FCCs compared with toddler classrooms. This pattern was similar on other ORCE scales, with toddler classroom mean scores either similar to or lower than FCCs and infants on all scales except Stimulation of Cognitive Development and Language Stimulation; for these, toddler classrooms had slightly higher mean scores than infant classrooms.

ORCE Training and Certification. Following the comprehensive self-study training program outlined by the ORCE developers, members of the Mathematica ORCE training team were certified in advance of the field observer training on the 6- , 15- , 24- , and 36-month age forms.[29] A team of seven prospective field observers attended a five-day comprehensive ORCE training led by the three certified Mathematica ORCE lead trainers. Prior to the training, trainees were required to review the coding definitions and exemplars provided for each of the behavior and qualitative rating scales, and completed a written quiz to assess their understanding of the material. The training included a series of video-based coding practice exercises and a live field observation in a center-based setting.[30] On the fifth day of training, trainees independently coded three certification videos, each consisting of one 44-minute ORCE cycle. The ORCE developers master coded the certification videos, which represent varying types of child care experiences in

[29] The ORCE developers made available to Mathematica the materials used to train and certify members of the ORCE team; these include age-specific coding manuals, video segments demonstrating low and high ends of the qualitative scales, 10- and 44-minute criterion video segments designed to expose trainees to a wide variety of child care settings and experiences, and accompanying master codes and written justifications for all provided videos. These master-coded videos serve as practice videos during training and as criterion videos for certifying field observers and monitoring ongoing reliability throughout the ORCE field period.

[30] Live practice observations were conducted in center-based settings to accommodate multiple observers.

center-based classrooms and FCCs serving infants and toddlers. To be certified as field observers, trainees were required to achieve exact agreement with the master codes on the Behavior Checklist items at a level of 70 percent or above, and with the qualitative rating scales at a level of 60 percent or above (NICHD Early Child Care Research Network 1996).

To monitor ongoing reliability, all certified field observers independently completed two further rounds of video-based reliability checks during the eight-week field period. Reliability checks occurred following the field observer's first week in the field, and again when the field observer had completed approximately half of his/her overall assignment. Field observers independently coded two 44-minute segments (master coded by developers) for each reliability check.

Table VI.2. Descriptive statistics for the ORCE scales, full sample

Scales	N	M (SD)	Minimum	Maximum	Cronbach alpha
Qualitative Rating Scales					
Overall Qualitative Rating	119	3.21 (.43)	1.79	3.95	.87
Sensitivity/Responsiveness to Distress[a]	81	3.55 (.64)	1.00	4.00	.93
Sensitivity/Responsiveness to Non-Distress	119	3.31 (.58)	1.10	4.00	.89
Lack of Intrusiveness	119	3.86 (.32)	2.13	4.00	.88
Lack of Detachment/Disengagement	119	3.69 (.49)	1.92	4.00	.90
Stimulation of Cognitive Development	119	2.12 (.64)	1.00	3.75	.81
Positive Regard for the Child	119	3.46 (.46)	2.25	4.00	.87
Lack of Negative Regard for the Child	119	3.98 (.10)	3.38	4.00	.74
Flatness of Affect	119	3.73 (.51)	1.75	4.00	.94
Fostering Exploration[b]	40	2.39 (.87)	1.00	4.00	.91
Positive Rating	119	2.91 (.51)	1.47	3.92	.87
Lack of Negative Rating	119	3.81 (.28)	2.78	4.00	.76
Behavior Checklist					
Language Stimulation	119	1.96 (.74)	0.63	4.13	.85
Positive Behavior Toward Child	119	0.62 (.54)	0.00	2.70	.92
Negative Behavior Toward Child	119	0.28 (.41)	0.00	1.53	.44
Child/Adult Ratio	118	3.25 (1.43)	1.00	8.50	--

Source: Q-CCIIT Fall 2012 Psychometric Field Test.

[a]Rated only if child distress occurred during the observation.

[b]For 36-month form only.

Table VI.3. Descriptive statistics for the ORCE scales, by child age and setting type

	Infant				Toddler				FCC			
Scales	N	M (SD)	Min.	Max.	N	M (SD)	Min.	Max.	N	M (SD)	Min.	Max.
Overall Qualitative Rating	37	3.26 (0.39)	1.97	3.85	41	3.14 (0.42)	1.90	3.77	41	3.24 (0.48)	1.79	3.95
Sensitivity/Responsiveness to Distress[a]	32	3.61 (0.50)	2.00	4.00	23	3.50 (0.79)	1.00	4.00	26	3.53 (0.68)	1.40	4.00
Sensitivity/Responsiveness to Non-Distress	37	3.35 (0.52)	2.17	4.00	41	3.24 (0.57)	1.10	4.00	41	3.34 (0.64)	1.50	4.00
Lack of Intrusiveness	37	3.88 (0.21)	3.13	4.00	41	3.87 (0.27)	2.75	4.00	41	3.82 (0.42)	2.13	4.00
Lack of Detachment/Disengagement	37	3.69 (0.48)	2.00	4.00	41	3.67 (0.55)	1.92	4.00	41	3.71 (0.45)	2.38	4.00
Stimulation of Cognitive Development	37	1.98 (0.66)	1.17	3.38	41	2.09 (0.58)	1.19	3.44	41	2.28 (0.65)	1.00	3.75
Positive Regard for the Child	37	3.53 (0.41)	2.56	4.00	41	3.41 (0.47)	2.33	4.00	41	3.44 (0.51)	2.25	4.00
Lack of Negative Regard for the Child	37	3.99 (0.06)	3.63	4.00	41	3.99 (0.03)	3.88	4.00	41	3.95 (0.15)	3.38	4.00
Lack of Flatness of Affect	37	3.77 (0.46)	1.94	4.00	41	3.69 (0.52)	1.75	4.00	41	3.74 (0.54)	2.00	4.00
Fostering Exploration[b]	--	--	--	--	19	2.30 (0.76)	1.00	3.75	20	2.46 (1.01)	1.00	4.00
Positive Rating	37	2.95 (0.45)	1.97	3.75	41	2.83 (0.48)	1.54	3.61	41	2.96 (0.58)	1.47	3.92
Lack of Negative Rating	37	3.83 (0.26)	2.89	4.00	41	3.80 (0.28)	2.83	4.00	41	3.80 (0.31)	2.78	4.00
Language Stimulation	37	1.81 (0.57)	0.72	3.15	41	1.89 (0.79)	0.66	3.75	41	2.17 (0.79)	0.63	4.13
Positive Behavior Toward Child	37	0.84 (0.63)	0.08	2.70	41	0.48 (0.41)	0.04	1.83	41	0.56 (0.51)	0.00	2.44
Negative Behavior Toward Child	37	0.50 (0.48)	0.00	1.53	41	0.10 (0.26)	0.00	1.50	41	0.27 (0.38)	0.00	1.39

Source: Q-CCIIT Fall 2012 Psychometric Field Test.

[a] Rated only if child distress occurred during the observation.

[b] For 36-month form only.

C. Validation measure: the ERS

The ERSs were developed to assess process quality in early child care settings. Process quality includes interactions among staff, children, parents, and other adults in a classroom. It also includes interactions children have with materials and activities in the environment, and features that support these interactions, such as arrangement of space and the classroom schedule. The Q-CCIIT study uses two versions of the scales relevant for child care settings for infants and toddlers: the ITERS-R for center-based programs caring for children from birth to 30 months of age; and the FCCERS-R for family child care settings for children from birth to 12 years. The ERSs define environment in a broad sense and guide the observer to assess the arrangement of space, materials, activities, supervision, interactions, and schedule.

ERS Observation Procedure. Each ERS observation began with the observer gathering key information from a primary caregiver, such as the total number of children currently enrolled in the class, the maximum number of children allowed to enroll, and whether there were any children with identified disabilities enrolled. After asking the teacher the preliminary questions, observers began their three-hour observation. Throughout its course, observers completed a score sheet with more than 30 items, each comprising a series of yes/no indicators used to arrive at a rating on a 7-point scale, with 1 representing inadequate, 3 minimal, 5 good, and 7 excellent.[31] At the end of the observation, observers asked primary caregivers scripted questions to evaluate any indicators they were not able to observe directly; for example, if art was not observed, observers asked questions about art materials.[32] Observers then used this additional information to complete their ratings.

Expected Relationship of the ERS and Q-CCIIT Measures. The ERSs measure both the environment and caregiver interactions, allowing examination of both convergent and discriminant validity of the Q-CCIIT. We expected moderate to strong correlations with the Listening and Talking and Interaction/Social scales in the ITERS-R in centers and FCCERS-R in FCCs (convergent validity), and weaker correlations with the scales that focus more on the environment, such as Space and Furnishings (discriminant validity). We conducted validation observations with the ITERS-R or the FCCERS-R in another 114 classrooms (ITERS-R in 65 centers and FCCERS-R in 49 FCCs).

The ITERS–R. The ITERS-R is an observational measure of center-based classroom quality designed for use in classrooms with children from birth to 30 months of age. It includes seven environmental subscales: (1) Space and Furnishings (5 items), (2) Personal Care Routines (6 items), (3) Listening and Talking (3 items), (4) Activities (10 items), (5) Interaction (4 items), (6) Program Structure (4 items), and (7) Parents and Staff (7 items). We used the first six subscales (32 items) in the Q-CCIIT validation observation, excluding the last subscale because these items rely heavily on staff reports rather than observations. The ITERS is used widely in early childhood research. Reported psychometric properties were good, with internal consistency for the total instrument reported at 0.93 (with some variation by individual subscale). The

[31] The ITERS-R had 32 items, and the FCCERS-R had 34 items.

[32] We trained observers to use the predetermined list of questions located at the bottom of the page in the measure whenever possible so as to avoid inadvertently asking questions in a leading manner. In addition, we trained them to allow approximately 20 minutes for asking caregivers these questions and wait until the caregiver was available to talk, even if that meant coming back during a more convenient time, such as at nap time.

authors reported finding evidence of predictive validity for the predecessor of the ITERS-R (the ITERS) and cited strong relations between the ITERS and the ITERS-R as evidence of concurrent validity (Harms et al. 2003). The ITERS has been used in programs serving children from culturally and linguistically diverse populations in the United States (for example, many state child care quality rating and improvement systems [QRIS] use or have used the ITERS).

The FCCERS–R. The FCCERS-R is an observational measure of child care quality designed for use in home-based settings with children ages birth to 12 years of age. Similar to the ITERS-R, the FCCERS-R also includes seven environmental subscales: (1) Space and Furnishings (6 items), (2) Personal Care Routines (6 items), (3) Listening and Talking (3 items), (4) Activities (11 items), (5) Interaction (4 items), (6) Program Structure (4 items), and (7) Parents and Staff (4 items). We used the first six subscales (34 items) in the Q-CCIIT validation observation. The FCCERS-R has been used widely in programs serving children from culturally and linguistically diverse populations in the United States (for example, many state child care QRIS use or have used the FCCERS-R or its predecessor, the Family Day Care Rating Scale [FDCRS]). Reported psychometric properties were good, with internal consistency for the total instrument reported at 0.90 (with some variation by individual subscale). No information was available on concurrent or predictive validity with child outcomes for infants and toddlers.[33]

The six FCCERS-R subscales are the same as the ITERS-R subscales, with some of the items tailored for family child care settings. We created subscale scores for each of them. Possible scores range from 1 to 7, with ratings of 1 indicating minimal quality, 3 moderate quality, 5 good practice, and 7 excellent quality (Harms et al. 2003).

Descriptives of the ERS Subscales in the Q-CCIIT Field Test. Tables VI.4 and VI.5 present the descriptive statistics for the ITERS-R and FCCERS-R for the full sample and the ITERS-R by child age.[34] In the areas of particular interest to Q-CCIIT—that is, the subscales assessing interaction—we had a wide range of scores on both the ITERS-R (1.3 to 7.0) and the FCCERS-R (1.0 – 7.0) with standard deviations greater than 1.4, suggesting that we met our goal of maximizing variation. The total score and activities and personal subscale scores have lower means and a more restricted range than hoped for on both the ITERS-R and the FCCERS, but that may be due to changes in the stringency of the scoring criteria[35] rather than a restricted range in the types of programs in the sample.

[33] A recent evaluation of Colorado's Qualistar QRIS program (Zellman et al. 2010) examined the correlation between the FDCRS and outcomes for children ages 2.5 through 5. (For approximately 40 homes and 120 children, results did not account for clustering.) Across three waves, no clear patterns emerged. At wave 1, no significant correlations were detected between the total score and the Woodcock-Johnson achievement subtests and the Child Behavior Inventory (CBI) subscales. At wave 2, with a smaller sample, significant correlations were found between the FDCRS total score and some CBI subscales. At wave 3, with ultimately about 20 FCCs (number of children not reported), only one significant correlation between the FDCRS total score and letter-word identification was evident.

[34] The FCCs were mixed age groups.

[35] For example, for naps, there is a new note forbidding the practice of swaddling, the use of cribs with sides that drop down, and the use of the ends of cribs as barriers to prevent the spread of disease (that is, cribs/cots must be separated by three feet of space even if they have solid ends).

Table VI.4. Child-adult ratio, total and subscale scores for the ITERS-R and FCCERS-R, full sample

	ITERS-R				FCCERS-R			
Subscale	N	Mean	SD	Range	N	Mean	SD	Range
Total	65	3.29	0.84	1.6 – 5.1	49	2.83	0.78	1.1 – 3.9
Listening and Talking	65	3.67	1.42	1.3 – 7.0	49	3.63	1.28	1.0 – 6.7
Social Interaction	64	4.33	1.52	1.3 – 7.0	49	4.30	1.71	1.0 – 7.0
Activities	65	3.12	0.99	1.3 – 5.3	49	2.45	0.80	1.1 - 3.8
Program Structure	64	3.55	1.38	1.5 – 7.0	49	3.34	1.44	1.0 – 6.7
Space and Furnishings	65	3.76	0.98	1.6 – 6.0	49	2.74	0.86	1.2 – 4.7
Personal Care	65	2.08	0.75	1.0 – 5.0	49	1.92	0.65	1.0 – 3.8
Child/Adult Ratio	65	3.94	1.79	1.5 – 14	48	5.01	2.53	2 - 14

Source: Q-CCIIT Fall 2012 Psychometric Field Test.

Note: New clarifications added to the ITERS-R and FCCERS-R in 2012 may have led to lower scores.

ITERS-R = Infant-Toddler Environment Rating Scale-Revised.

Table VI.5 presents the descriptive information for the ITERS-R subscales by child age. The Activities subscale score was significant lower in classrooms serving infants than in those serving toddlers. The subscale scores also were lower in classrooms serving infants for the Listening and Talking, Program Structure, Space and Furnishings subscales, and the total score, although not statistically significant.

ERS Training and Certification. Following the training program outlined by the ITERS-R and FCCERS-R developers, a team of three lead trainers conducted eight days of training—four per ERS measure. We trained seven observers across both measures. Prior to the training, trainees were required to review the ITERS-R and FCCERS-R measures, including the definitions and "Notes for Clarification," and complete a written quiz to assess their understanding of the material. We addressed any incorrect responses or misunderstandings during the in-person training.

The first two days of each training consisted of classroom instruction, which included a thorough review of the measure, completion of training exercises, and numerous opportunities to practice coding using videos and training workbooks produced by the measure developers. On the second two days of training for each measure, observers conducted classroom observations with a gold standard observer.[36] To be considered reliable, the field staff trainees were required to match within one point of the consensus score on the seven-point scale for 80 percent of the items within the measure; scores from the two ITERS-R visits were averaged for each measure, as were the scores from the two FCCERS-R visits.[37]

[36] All gold standard observers demonstrated 85 percent reliability with the ITERS /FCCERS lead trainers prior to the main study training.

[37] At the end of each observation, the gold standard observers discussed their scores and any areas of difficulty, and established final consensus scores. Reliability was measured against the consensus score based on the recommendation of the measure developers (D. Cryer, personal communication, June 14, 2012).

Table VI.5. Child-adult ratio and subscale scores for the ITERS-R, by child age

	Infant				Toddler			
Subscale	Mean	SD	Min	Max	Mean	SD	Min	Max
Listening and Talking	3.45	1.22	1.33	5	3.86	1.60	1.67	7
Social Interaction	4.39	1.45	1.25	6.75	4.27	1.63	1.75	7
Activities	2.77**	0.88	1.25	4.86	3.39	0.96	2	5.22
Program Structure	3.38	1.41	1	6	3.67	1.33	1	7
Space and Furnishings	3.71	1.07	1.6	5.8	3.79	0.90	2.2	6
Personal Care	2.16	0.66	1	3.67	1.98	0.83	1	5
Total	3.17	0.81	1.62	4.89	3.37	0.86	1.8	5.07
Child/Adult Ratio	3.52†	1.30	1.5	6	4.30	2.13	2.67	14
Sample Size	**32**				**31–32**			

Source: Q-CCIIT Fall 2012 Psychometric Field Test.
Note: Item ratings range from 1 to 7.
†p< .10.
**p<.01.

To monitor ongoing reliability, a gold standard observer observed all certified field observers again during the first two weeks of the field period. These quality assurance visits followed the same procedures used at training.

D. Sample

Classroom Roster and Caregiver Self-Administered Questionnaire (SAQ). In each observed classroom, we collected information about the classroom and eligible caregivers. The primary caregiver/teacher answered questions on the Classroom Roster about the children in the classroom (for example, gender, age, languages spoken in the children's homes). We asked the caregivers to report on the dates of birth for all children in the classroom *present on the day that we observed*.[38] This information was the primary source for categorizing center-based classrooms as "infant" or "toddler." We received rosters for 99 percent of the classrooms. We asked all eligible[39] caregivers in the observed classrooms to report information about their background characteristics (such as education, experience, and depressive symptoms). We received at least one caregiver questionnaire from 98 percent of the classrooms. We expected these classroom and caregiver characteristics to be associated with the quality in the classroom and assessed the validity of the Q-CCIIT measure by examining the associations with these characteristics.

Classroom and Caregiver Characteristics. For both FCCs and center-based classrooms, we also explored the validity of the Q-CCIIT with caregiver experience and structural features,

[38] We detected some error in this reporting, as caregivers sometimes reported more dates of birth than number of children noted by the Q-CCIIT observers during their Q-CCIIT cycle ratings.

[39] Eligible caregivers were defined as paid caregivers who were in the room daily for four or more hours.

such as child-to-staff ratios. We collected information on these measures using the Caregiver SAQ and Classroom Roster. A copy of these instruments along with frequencies for each question can be found in Appendix A.

The Q-CCIIT sample included 403 classrooms.[40] Of these classrooms, 85 were affiliated with Early Head Start programs (6 of these were FCCs), and 23 were state-funded child care centers. As depicted in Table VI.6, a language other than English was spoken in 38 percent of the classrooms. Across all of the students in all of the classrooms, 31 percent of the children were Dual Language Learners (DLL), and 15 percent had an Individual Family Service Plan (IFSP).

The majority of the caregiver respondents were full-time employees (81 percent), and 43 percent of the caregivers reported working with part-time staff. A total of 94 percent of the respondents described themselves as a Lead Teacher, Assistant Teacher, or Teacher, and 49 percent had an associate's or more advanced degree; the majority had college coursework in Early Childhood Education or Child Development. Most had been in their current classroom for at least one year, with 20 percent in their current classroom for five or more years. Twenty-two percent of caregivers spoke a language other than English (often in addition to English).

FCCs were different from center-based classrooms in the age of the children in the class (fewer infants). There were also fewer FCC caregivers with an associate's degree or higher, and fewer FCC caregivers had coursework in Early Childhood Education or Child Development. FCC caregivers were more likely to have been in their current setting and classroom for five or more years than center-based caregivers, although they were less likely to describe their position as full time.

In the next chapter, we describe results of the analyses conducted for the psychometric field test.

[40] Note: we included in subsequent analyses 400 of the classrooms with at least five valid observation cycles (that is, observation cycles that lasted five minutes or longer).

Table VI.6. Sample characteristics for the overall sample and by program type and age group

	Full Sample	Classroom Type		
		Infant	Toddler	FCC
	Percentage	Percentage	Percentage	Percentage
Children in Classrooms Characteristics				
Children younger than 18 months[a]	42.68	93.24	9.62	29.19
Children 18–36 months	47.43	6.11	83.05	46.53
Children who are DLL	31%	27%	34%	31%
Children with IFSPs	15.03	13.78	14.38	17.27
Sample Size	**2,158**	**729**	**1,065**	**364**
Classroom Characteristics				
Language other than (or in addition to) English spoken in the classroom	38%	43%	37%	33%
Sample Size	**403**	**138**	**155**	**110**
Caregiver Characteristics				
Full time	81%	82%	85%	71%
CDA certification	35%	35%	37%	24%
Highest degree is high school	18%	19%	16%	18%
Some college courses (no degree)	32%	27%	32%	42%
Associate's degree or higher	49%	52%	51%	39%
One or more college course in Early Childhood Education	62%	65%	66%	48%
One or more college course in Child Development	61%	66%	64%	47%
One or more college course in Infant Development	50%	54%	53%	39%
At setting:				
Less than a year	20%	27%	19%	13%
1–2.5 years	17%	18%	19%	14%
2.5–5.5 years	22%	20%	25%	21%
More than 5.5 years	38%	33%	36%	50%
In classroom:				
Less than a year	36%	47%	38%	16%
1–2.5 years	20%	18%	26%	13%
2.5–5.5 years	19%	16%	18%	23%
More than 5.5 years	20%	15%	13%	40%
Female	97%	98%	98%	92%
Speak a language other than English	22%	23%	20%	25%
Hispanic	27%	23%	27%	32%
White, non-Hispanic	39%	42%	40%	35%
African American, non-Hispanic	28%	28%	28%	30%
Other	3%	4%	4%	>0%
Sample Size	**972**	**359**	**397**	**216**

Source: Q-CCIIT Fall 2012 Psychometric Field Test Data Collection based on caregiver responses to a classroom roster and a self-administered questionnaire.

Note: Additional information about item level medians can be found in Appendix A. Denominators include cases for which the variable was missing.

[a] These estimates are based on the children with valid ages between 0 and 36 months according to the caregiver-completed classroom roster. In the full sample, 181 children in infant classrooms, 306 children in toddler classrooms, and 287 children in FCCs had missing or invalid ages (less than 0 or greater than 36 months) for participation in the Q-CCIIT observations. All groups reported children older than 36 months: Infant classrooms = one child, Toddler classrooms = 88 children, and FCCs = 170 children.

This page has been left blank for double-sided copying.

VII. DATA ANALYSIS AND FINDINGS FOR THE PSYCHOMETRIC FIELD TEST

We conducted psychometric analysis with the field test data to evaluate item functioning and assess the reliability and validity of the Q-CCIIT measure. The goals of our analysis were to (1) assess reliability and stability of the measure, (2) assess the construct and concurrent convergent validity evidence, and (3) examine invariance across subgroups (that is, look for the absence of any differential item functioning [DIF]). We used approaches that draw on classical test theory, generalizability theory, and item-response theory (IRT).

Specifically, our analyses included the following:

- Examining variance in item descriptives, both overall and by subgroups (program type and child age), as well as the activity context for scores
- Estimating reliability of scores, including inter-rater reliability, test-retest reliability (temporal stability), and internal consistency reliability (Cronbach alpha), as well as reliability from the perspective of generalizability theory (G-study) and IRT
- Conducting confirmatory factor analysis (CFA) across all classrooms, as well as by subgroups (program type and child age)
- Examining item difficulties, level of discrimination among classrooms, distribution of performance relative to difficulty of implementing practices, and fit statistics across all classrooms, as well as conducting DIF analysis by program type and child age
- Examining evidence of concurrent validity across all classrooms, as well as conducting subgroup analyses by program type and child age

A. Observation context for scores

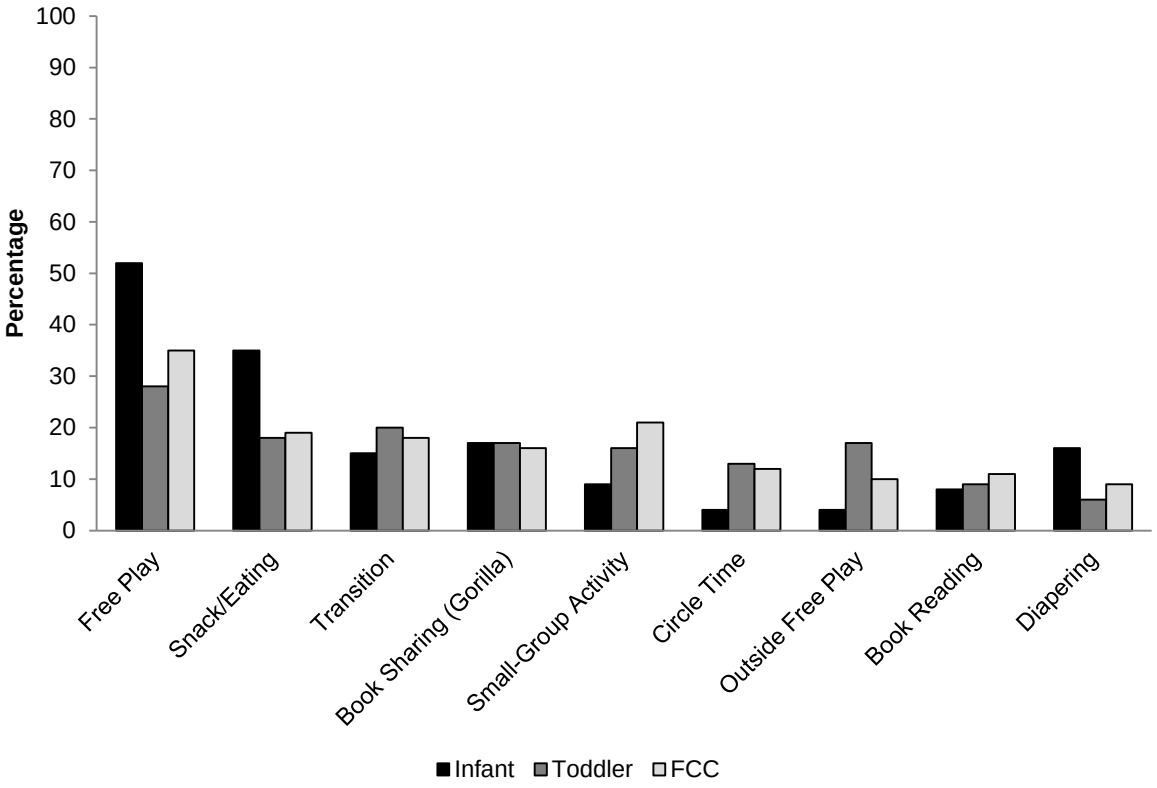
The field test observations took place during a variety of caregiving activities. During each 10-minute observation cycle, observers noted the type of interaction occurring and could note more than one type within a cycle. The majority of classrooms had six cycles, with one observation cycle representing approximately 17 percent of the cycles in a given classroom. Across the entire sample (N = 400), free play was the most commonly occurring activity (Appendix C, Table C.1), with 38 percent of the observation cycles[41] including free play (SD = 0.26). This was followed by eating or feeding (M = 0.24, SD = 0.18); transitions between activities (M = 0.18, SD = 0.19); book sharing of *Good Night, Gorilla* (M = 0.17, SD = 0.05); and small-group, teacher-directed activities (M = 0.15, SD = 0.18).

These overall results were largely consistent across classroom types (Figure VII.1), with infant (N = 136), toddler (N = 154), and family child care home (FCC) (N = 110) classrooms all spending a majority of the observed time in a combination of free play, eating and feeding, transitions between activities, and the requested book sharing of *Good Night, Gorilla* (Appendix C, Table C.1). Additionally, FCC classrooms spent approximately 21 percent of the observed time in small-group, teacher-directed activities, compared to only 9 percent of infant classrooms

[41] The percentage of observation cycles during which each type of interaction was observed was calculated as the mean across all valid observation cycles (cycles lasting 5 minutes or longer). On average, there were a total of six observation cycles (10 minutes each) during each visit.

and 16 percent of toddler classrooms. Infant classrooms were more likely to include a cycle with diapering or toileting than other settings. Compared with infant classrooms, toddler classrooms were more likely to include a cycle with outside free play.[42]

Figure VII.1. Observation context, by classroom type



Source: Q-CCIIT Fall 2012 Psychometric Field Test Data.

Note: Observed classrooms: 136 infant, 154 toddler, and 110 FCC classrooms.

B. Item level descriptive statistics

For the field test, we examined item-level descriptive statistics for the Q-CCIIT measure for the full sample and by setting type (center-based program versus FCC) and child age (infant versus toddler), as we did in the pilot test. Appendix C presents these item-level statistics for all items in the order in which they were rated on the Q-CCIIT form. Across all positive items (that is, items addressing support for children's development), there was adequate variance and no indication of a ceiling or floor problem for the full sample. Mean scores on items ranged from 3 to 5 (out of 7), with standard deviations from 1 to 2. The range for most items was from 1 to 7. For infant classrooms, a skew was evident for the Support for Cognitive Development items, with means for some items between 2 and 3 and standard deviations of 1.3 to 1.6 (Appendix C, Table C.9). As expected, the Areas of Concern items had limited variance, with no or few concerning behaviors in most classrooms.

[42] We did not test for significance for these comparisons.

Support for Social-Emotional Development. Across all classrooms and programs (*N* = 266–400),[43] caregivers' scores on social-emotional development items that involved responding to children's cues (distress, social cues, and emotional cues) and building positive relationships were, on average, between 4.5 and 5 (Appendix C, Table C.8). For the four Support for Social-Emotional Development items rated across the visit[44]—classroom limits and management, responsive routines, sense of belonging, and supervising or joining in play or activities—scores for the full sample ranged between 3.65 (*SD* = 1.29) and 4.70 (*SD* = 1.6).

When examining the results by classroom type, scores were highest in toddler classrooms (*N* = 94 – 154) for almost all of the social-emotional development items, with two exceptions: infant classrooms had the highest mean ratings on items regarding responding to distress and building positive relationships (Appendix C, Table C.7). Overall, FCC classrooms tended to score lower than infant classrooms.

Support for Cognitive Development. On the cognitive development items rated within each 10-minute cycle—supporting object exploration, scaffolding problem solving, and supporting peer interaction and play—the means for the full sample (N = 372–399) were 3.84, 3.06, and 3.33, respectively (Appendix C, Table C.10). For the four cognitive development items rated across the visit, the mean scores were highest for giving choices (M = 3.58), explicit teaching (M = 3.31), and extending pretend play (M = 3.14) (Appendix C, Table C.10).

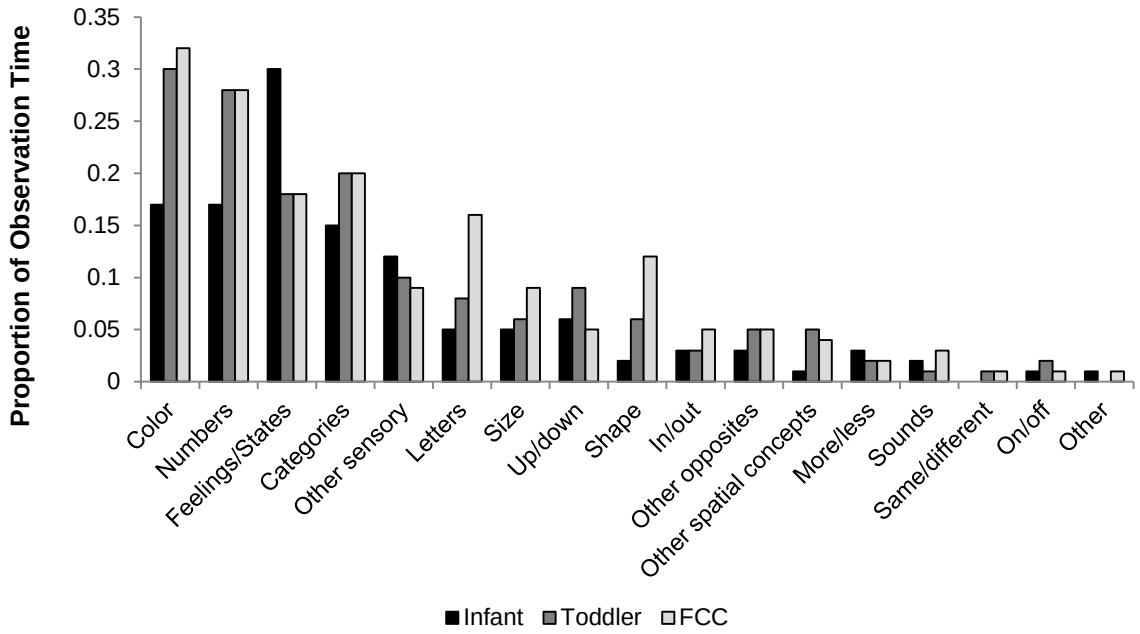
This pattern of scores was consistent across infant, toddler, and FCC classrooms (Appendix C, Table C.9), with toddler classrooms consistently scoring highest on Support for Cognitive Development items, infant classrooms consistently scoring lowest, and FCC classrooms scoring in the middle.

Concept Development. Observers were asked to indicate whether or not concepts were introduced in each observation cycle. Across the full sample, an average of five unique concepts (*M* = 5.28, *SD* = 2.59) were presented during any single observation (Appendix C, Table C.5). The most commonly observed concepts included color, number, feelings or states, categories (animals, furniture, etc.), and other sensory concepts, such as textures or how things smell or taste. Letters, size, common opposites (up/down, in/out, same/different, more/less), shapes, sounds, and other spatial concepts were more rarely observed (Figure VII.2). Looking across classroom type (Appendix C, Table C.5), FCCs and toddler classrooms provided more frequent opportunities for concept development than infant classrooms. The mean number of unique concepts presented in classrooms was greatest in FCCs.

[43] Some items have missing data because of the "not applicable" (NA) option. For example, Response to Distress is not coded if there is no distress.

[44] Across-the-visit items were completed at the end of the classroom observation; observers reviewed their notes and score sheet to rate these items.

Figure VII.2. Use of concepts, by classroom type



Source: Q-CCIIT Fall 2012 Psychometric Field Test Data.

Note: Observed classrooms: 136 infant, 154 toddler, and 110 FCC classrooms.

Support for Language and Literacy Development. For items rated within each 10-minute observation cycle, support for children's language and literacy development was greater during book sharing than other observed times, with scores, on average, between 4.5 and 5 for variety of words and engaging children in books (Appendix C, Table C.12). During times other than book sharing, scores for items rated within the 10-minute segments ranged between 3 and 4, with the highest mean score for caregiver use of varied vocabulary (M = 3.98, SD = 1.19) and the lowest mean score for extending children's language use (M = 3.35, SD = 1.20). For the three language items rated across the visit—features of talk, talk about things not present, and positive attitude toward books—caregivers received the highest score for features of talk (M = 4.94, SD = 1.45) and the lowest for talk about things not present (M = 3.26, SD = 1.49) (Appendix C, Table C.12).

When examining scores by classroom types, caregivers in toddler classrooms received the highest language scores, while caregivers in FCC classrooms received higher scores than those in infant classrooms on all language and book-sharing items (Appendix C, Table C.11). In general, infant classrooms provided less language than toddler and FCC classrooms. Consistent with the items rated within the 10-minute segments, the across-the-visit language items were rated highest in toddler classrooms and lowest in infant classrooms (Appendix C, Table C.11). Toddler classrooms experienced more variation in vocabulary and types of talk (features of talk) than FCC classrooms, but otherwise were similar.

C. Scoring approach

We tested different approaches to creating Q-CCIIT scales. As mentioned above, the Q-CCIIT measure includes checklists, frequency ratings, and rubric ratings. Some items are repeated measures (collected for each cycle), and some are collected only at a single time point

(usually at the end of the observation or across the visit). For items collected through the 10-minute observation cycles, we calculated the mean across all valid cycles [45] and then created a mean score (including relevant items rated at the end of observation period) for each of the scales. For Concept Development, we calculated the diversity of concepts (that is, the number of different concepts presented in at least one of the observation cycles), the mean number of concepts during each cycle, and the frequency of discussion of concepts (the number of cycles in which any concept was presented). The mean number of concepts per cycle and the frequency of discussion of concepts were not related to measures of language or cognition and so were dropped from further analysis. We included the diversity of concepts in the measurement of Support for Cognitive Development. We also checked for differences in scores for different activity contexts to determine how critical it is to include these contexts in Q-CCIIT observations. [46]

In addition to the scoring approaches examined with the pilot test data, we tested additional approaches to scoring with the field test data. For example, we examined if the individual Types of Talk (for example, descriptive talk, explanations and reasoning) should be included in the Support for Language and Literacy Development and Support for Cognitive Development scales or if we should create a separate single score from those items.

We based all further analyses on mean scores for the scales identified in the exploratory factor analyses and supported in our confirmatory factor analyses (see Table VII.9).

D. Descriptive statistics for scale scores

We examined the scale scores by key subgroups of interest outlined in the proposed sampling design, such as classrooms with high and low proportions of children who are dual-language learners (DLLs). Consistent with the findings from the item-level descriptive statistics, classrooms serving infants tended to have lower mean scores than FCCs and classrooms serving toddlers, particularly in Support for Cognitive Development and Support for Language and Literacy Development (Table VII.1).

[45] Each cycle has to be five minutes or longer to be considered valid; we limited analyses to classrooms with at least five observation cycles of five minutes or more. We excluded three classrooms from the analysis due to an insufficient number of valid observation cycles.

[46] We noted differences in the quality of support for language and literacy development during the book-sharing cycles (stronger scores during book sharing). We did not find differences in scores for the meal or feeding cycles, except for the use of NA. NA is allowed for some items, such as "supporting object exploration" during meals or feeding.

Table VII.1. Q-CCIIT scales for the overall sample, and by child age and program type

	Total				Infant				Toddler				FCC				Centers			
	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
Support for Social-Emotional Development	4.53	1.10	1.31	6.89	4.44*	1.12	1.52	6.89	4.76***	0.99	1.54	6.69	4.31	1.19	1.31	6.88	4.61*	1.06	1.52	6.89
Support for Cognitive Development	3.45	1.02	1.14	6.31	3.05***	0.99	1.14	5.50	3.72	0.91	1.54	5.83	3.55***	1.05	1.18	6.31	3.40	1.00	1.14	5.83
Support for Language and Literacy Development	4.06	0.99	1.47	6.75	3.70***	0.95	1.65	5.82	4.34	0.86	1.50	6.12	4.12***	1.07	1.47	6.75	4.04	0.96	1.50	6.12
Sample size	**400**				**136**				**154**				**110**				**290**			

Source: Q-CCIIT Fall 2012 Psychometric Field Test.

Note: Scales for social-emotional, language, and cognitive development are based on exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) results.

The results of the significance test for infant vs. toddler are indicated on the infant mean. The results for toddler vs. FCC are on the toddler mean. The results for FCC vs. infant are on the FCC mean. The results for the center vs. FCC are on the Center mean.

*p<.05; **p<.01; ***p<.001.

E. Classic psychometric reliability analysis

The Q-CCIIT demonstrated strong internal consistency for the positive scales across all subgroups and all approaches to measurement. The Areas of Concern had limited variance and demonstrated weaker reliability. Using classical test theory approaches, we computed the coefficient alpha for each of the Q-CCIIT scales to assess internal consistency reliability (Table VII.2). For the full sample, the internal consistency reliability estimates were strong for the Support for Social-Emotional Development (0.93), Support for Language and Literacy Development (0.92), and Support for Cognitive Development (0.89) scales.[47] The reliability estimates were slightly weaker for Areas of Concern with Extreme Concern (0.83) and without Extreme Concern (0.80). For the Support for Social-Emotional Development, Support for Language and Literacy Development, and Support for Cognitive Development scales, the reliability estimates were similar between FCCs and centers. The reliability estimates for Areas of Concern were stronger in FCCs than centers due to greater variation in FCCs. FCCs had more frequent use of media, child restriction, interactions with verbal harshness, and chaos, as well as more evidence of unsafe or unhealthy practices.

The reliability estimates were similar for infant and toddler classrooms for Support for Social-Emotional Development and Support for Language and Literacy Development. Infant classrooms had stronger reliability than toddler classrooms on Support for Cognitive Development, while toddler classrooms had stronger reliability on Areas of Concern than infant classrooms.

Table VII.2. Internal reliability estimates (Cronbach alpha) of the Q-CCIIT scales for the overall sample, and by child age and program type

		Child age		Program type	
Subscales	Total	Infant	Toddler	Center	FCC
Support for Social-Emotional Development	0.93	0.93	0.93	0.93	0.93
Support for Language and Literacy Development	0.92	0.91	0.89	0.91	0.93
Support for Cognitive Development	0.89	0.91	0.84	0.88	0.90
Areas of Concern (with Extreme Concern)	0.83	0.62	0.70	0.67	0.84
Areas of Concern (without Extreme Concern)	0.80	0.63	0.75	0.70	0.81

Source: Q-CCIIT Fall 2012 Psychometric Field Test.

Note: Scales for social-emotional, language, and cognitive development are based on EFA and CFA results.

[47] The reliability results presented here are based on the final measure derived from exploratory (EFA) and confirmatory factor analysis (CFA).

As we did in the pilot test, we examined variation in the use of rating categories in an item and the item-total correlations[48] for each item. The results indicated that each of the items discriminates well among classrooms in measuring the specific construct. Item scores ranged from 1 to 7 on most items. Item-to-total correlations ranged from 0.60 to 0.82 across the scales, with the range on each scale being similar: Support for Social-Emotional Development ranged from 0.72 (response to distress) to 0.85 (response to emotional cues); Support for Language and Literacy Development ranged from 0.67 (positive attitude toward books) to 0.82 (conversational turn-taking), and Support for Cognitive Development ranged from 0.60 (diversity of concepts) to 0.80 (supporting object exploration).

We examined the ordering of categories of item rubrics and the fit of the steps from one category to the next in the Rasch IRT model. The results confirmed the ordering of the item rating categories on the Q-CCIIT scales, with the Rasch-Thurstone[49] thresholds increasing for categories 1 to 7 on each scale. The estimated discrimination for each category ranged from 0.70 to 1.21 for categories on the Support for Social-Emotional Development; from 0.81 to 1.13 for categories on Support for Language and Literacy Development; and from 0.89 to 1.10 for categories on Support for Cognitive Development.

Summary. The Q-CCIIT measure demonstrated acceptable internal consistency reliability, with strong reliability for the positive scales and weaker, though sufficient, reliability for the Areas of Concern (with the exception of the infant age group, which lacked variability). Variation in the use of rating categories in an item and the item-total correlations for each item indicated that items discriminate well among classrooms in measuring the specific construct. Rasch analysis indicated that the item rating categories are ordered correctly, with increasing categories indicating higher quality skill.

F. Test-retest reliability (temporal stability)

For 62 settings (32 center-based and 30 FCCs), observers returned for a second visit within one week of the first visit and completed a second Q-CCIIT observation. Although the observer was the same on the second visit, the combination of caregivers and children in the classroom could change from one visit to the next.

We examined the temporal stability (one-week test-retest reliability) by estimating Pearson correlation coefficients both for the scale scores and individual items. We conducted these analyses for the 62 settings, and by subgroup. We also examined the mean of the absolute value of the differences between the scores on the first and second visits for the overall scales and the individual items.

The results for the overall sample for infant classrooms and FCCs indicate adequate stability of the measure across days (Table VII.3). There was greater fluctuation in caregiver responsiveness and support in the toddler classrooms. It is not clear if this was due to differences

[48] Negative item-total correlations indicate that something should be reverse coded. A low item-total correlation indicates that the item is not providing strong information about the specific construct. It may contain too much information from a different dimension (adding multidimensionality).

[49] A Thurstone model uses a cumulative normal function, while Rasch incorporates a person parameter using a logistic function to examine whether rating categories are ordered correctly.

in the caregivers present across days, different child attendance across days, or differences in the children's states on different days. Examination of item-level correlations in the Support for Social-Emotional Development scale indicated a lack of significant correlation for response to distress, response to social cues, building a positive relationship, and responsive routines (Appendix D, Table D.1). In both the toddler classrooms and the FCCs, scores on the literacy items were not significantly correlated across days (Appendix D, Table D.2). On both days, the caregivers were asked to read *Good Night, Gorilla.* We hypothesize that the book either was read by a different caregiver or the caregiver read to a child of a different age (particularly in the case of the FCCs), conducted a re-reading in a different way, or read additional books with different levels of quality than were evident on the first day.

Summary. The Q-CCIIT measure demonstrated adequate temporal stability ($r \geq 0.70$) for the overall sample and the infant classrooms and FCCs. Estimates were lower in toddler classrooms on the Support for Language and Literacy scale ($r = 0.56$) and the Support for Social-Emotional Development scale ($r = 0.43$). The latter scale's lower reliability was surprising, since we would expect caregivers to be more consistent across days in Support for Social-Emotional Development. In addition, the book-sharing items within Support for Language and Literacy were not correlated across days; as the "easiest" items, they affected the scale correlation. These findings underscore the transactional nature of caregiver-toddler interactions: caregivers are not providing consistent support across days in classrooms in which toddler temperament can be more variable. These findings also suggest implications for high-stakes observations: children seem to drive more of the interactions in toddler than infant settings. With more variability in toddler mood/emotion regulation, caregivers may be more challenged in responding consistently to children.

Table VII.3. Test-retest correlations for scale scores overall, and by setting type

Scales	Total	Infant	Toddler	FCC	Center
N	61–62	18	14	30	32
Support for Social-Emotional Development	0.76**	0.85**	0.43	0.80**	0.73***
Support for Language and Literacy Development	0.79**	0.88**	0.56*	0.74**	0.82***
Support for Cognitive Development	0.79**	0.88**	0.88**	0.69**	0.89***
Chaos	0.76**	—	—	—	—
Extreme Areas of Concern	0.71**	—	—	—	—

Source: Q-CCIIT Fall 2012 Psychometric Field Test.

Note: With little to no variance within subgroups in the negative scales, we did not estimate the subgroup correlations for Chaos and Extreme Areas of Concern.

*p<.05; **p<.01.

G. Inter-rater reliability

In addition to our certification requirements both for video-based reliability with developer scores and field reliability with a gold standard observer, we also assessed rater reliability throughout the field period, using a combination of observations with gold standard observers and paired observations (other observers). We evaluated reliability by examining rater adjacent

agreement, correlations of the scale scores, Kappa coefficients, and difference scores between raters on the item-level and scale scores. We also estimated the amount of variance in the scores attributable to the observer in the generalizability analysis (G-studies).

Fifty-two paired field observations were conducted over a 10-week period, with 41 taking place in centers (17 in infant classrooms, 24 in toddler classrooms), and 11 in FCCs. Across the data collection period, 23 occurred early in the field period (weeks 1 through 3): 5 infant classrooms, 9 toddler classrooms, and 9 FCCs. Sixteen occurred mid-way through the field period (weeks 4 through 7): 8 infant, 6 toddler, and 2 FCC. Thirteen occurred late in the field period (weeks 8 through 10): 4 infant and 9 toddler. [50]

We examined adjacent agreement (within one point) for the ratings and exact agreement for checklists and Areas of Concern. For the items coded during each cycle, we averaged the percentage agreement across all cycles. Table VII.4a presents the average item-level agreement across pairs overall and by setting type. Across the entire rating form (all item ratings and checklists in the rating form except for types of talk), the level of agreement was 87 percent overall, ranging from 84 percent for FCCs to 89 percent for center-based infant classrooms. All scales had an agreement of 84 percent or higher: 88 percent for Support for Social-Emotional Development, 88 percent for Support for Language and Literacy Development, and 84 percent for Support for Cognitive Development. For overall setting description, the agreement was 90 percent or higher: 92 percent for Environment, 93 percent for Areas of Concern, and 98 percent for Extreme Areas of Concern. Agreement was high across setting types, though somewhat weaker in FCCs. In this sample of classrooms, no Extreme Areas of Concern in infant or toddler settings occurred.

Table VII.4a. Q-CCIIT Inter-rater field reliability, mean percentage item agreement, overall and by setting type

Scales	Total	Infant	Toddler	FCC
Total	87.4	88.5	88.2	83.9
Support for Social-Emotional Development	88.2	89.0	89.1	85.2
Support for Language and Literacy Development	88.3	92.4	87.1	84.5
Support for Cognitive Development	83.9	89.0	83.3	77.3
Areas of Concern	92.8	94.1	96.1	83.5
Extreme Areas of Concern	98.1	100	100	90.9
Sample Size	**52**	**17**	**24**	**11**

Source: Q-CCIIT Fall 2012 Psychometric Field Test Data.

[50] A greater proportion of Q-CCIIT inter-rater reliability visits occurred during the early field period because one week had passed after training. Paired observations provided a check that field staff were maintaining their skills. Additionally, gold standard reliability checks occurred during week 3 to ensure against observer drift. A greater proportion of FCC reliability visits occurred early in the field period to limit the number of observers in the setting, given that the ORCE validation observations occurred mid-way through the field period. In FCC homes, we limited the total number of observers to two field staff. To complete the required number of paired observations, we had to balance the Q-CCIIT inter-rater reliability and the Q-CCIIT-ORCE and Q-CCIIT-FCCERS co-observations. One paired observation occurred at the end of week 12.

We also examined the agreement on the scale scores. Overall, the scale score reliability was high, with agreement exceeding 85 percent, strong correlations (that is greater than 0.70), and difference scores of less than one-half of a point. All scales had adjacent agreement of 87 percent or higher: 89 percent for Support for Social-Emotional Development, 87 percent for Support for Language and Literacy Development, and 90 percent for Support for Cognitive Development (Table VII.4b).

Correlations between observers' scores were 0.75 or higher: 0.82 Support for Social-Emotional Development, 0.75 for Support for Language and Literacy Development, and 0.78 for Support for Cognitive Development (Table VII.5a). Difference scores averaged across paired observations were 0.48 or smaller, ranging from 0.42 for Support for Social-Emotional Development to 0.48 for Support for Cognitive Development (Table VII.5b).

The evidence of rater reliability on the resulting scale score varies somewhat by setting type. Across scores, many settings showed strong reliability, with percentage agreement of 90 percent or greater, correlations of 0.76 or higher, and difference scores at or below one-half point. While some scales had inter-rater reliability below the criteria set by the study team (80 percent agreement, a 0.70 correlation, and less than one-half point difference), the inter-rater reliability estimates differed across the three criteria, demonstrating adequate reliability with at least one of the methods. The sample sizes by setting type ranged from 11 to 24, so a difference between a few observers can affect the correlation considerably, compared to what we might find in a larger sample.

By setting, the following scales had lower rater reliability on at least one of the criteria:

- For infant classrooms, Support for Social-Emotional Development was lower than criteria but close to typical high cut-points, with 77 percent of classrooms with final scale scores agreeing within one point, a moderate correlation (r = .66), and a difference score close to one-half of a point (0.51).
- For toddler classrooms, Support for Language and Literacy Development was somewhat lower, with a correlation of 0.43 between observers' scores, but the percentage agreement was 83 percent, and the difference score was close to one-half (0.51). While the correlation for the Support for Cognitive Development score was somewhat lower, at 0.60, the percentage agreement was high (92 percent), with a mean difference score of 0.46.
- For FCCs, Support for Language and Literacy Development and Support for Cognitive Development had lower percentage agreement (73 and 82 percent, respectively) and higher difference scores (in the 0.60s). However, the correlations were 0.76 and 0.79.

Table VII.4b. Q-CCIIT inter-rater field reliability, mean percentage agreement on scale scores, overall and by setting type

Scales	Total	Infant	Toddler	FCC
Support for Social-Emotional Development	88.5	76.5	95.8	90.9
Support for Language and Literacy Development	86.5	100.0	83.3	72.7
Support for Cognitive Development	90.4	94.1	91.7	81.8
Sample Size	**52**	**17**	**24**	**11**

Source: Q-CCIIT Fall 2012 Psychometric Field Test Data.

Table VII.5a. Q-CCIIT scale score correlations between observers

Scales	Total	Infant	Toddler	FCC
Support for Social-Emotional Development	0.82	0.66	0.79	0.93
Support for Language and Literacy Development	0.75	0.88	0.43	0.76
Support for Cognitive Development	0.78	0.86	0.60	0.79
Sample Size	**52**	**17**	**24**	**11**

Source: Q-CCIIT Fall 2012 Psychometric Field Test Data.

Table VII.5b. Q-CCIIT average difference on scale scores between observers

Scales	Total	Infant	Toddler	FCC
Support for Social-Emotional Development	0.42	0.51	0.36	0.38
Support for Language and Literacy Development	0.46	0.29	0.51	0.62
Support for Cognitive Development	0.48	0.38	0.46	0.66
Sample Size	**52**	**17**	**24**	**11**

Source: Q-CCIIT Fall 2012 Psychometric Field Test Data.

We also estimated weighted Kappas. Cohen's Kappas examine exact agreement,[51] and weighted Kappas also take into account that observers may assign a particular rating by chance. However, weighted Kappas are affected by the prevalence of behaviors. When something is rare, a low Kappa may not reflect the level of agreement. Kappas in the range of 0.41 to 0.60 are considered to reflect moderate agreement, 0.61 to 0.80 substantial agreement, and greater than 0.81 almost perfect agreement (Sim and Wright 2005). For the total Q-CCIIT (all item ratings and checklists in the rating form except for types of talk), the weighted Kappa was 0.58, ranging from 0.54 for toddlers and FCCs to 0.58 for infants (Table VII.6). The scales showed a mean weighted Kappa of 0.61 for Support for Social-Emotional and Cognitive Development and 0.54 for Support for Language and Literacy Development. Kappas were lower for Areas of Concern. Agreement differed somewhat by setting type but the weighted Kappas usually were within a

[51] We rounded cycle item mean scores to the nearest whole number in estimating the Kappa for the item.

0.10 range of each other. Weighted Kappas were lower (that is below 0.50) for toddlers on Support for Language and Literacy Development and for FCCs on Support for Cognitive Development. Areas of concern had a wider range (0.57 for infants to 0.75 for toddlers), perhaps reflecting prevalence of these behaviors for infants.

Table VII.6. Q-CCIIT inter-rater reliability, weighted Kappas for scales, overall and by setting type

	Total	Infant	Toddler	FCC
Total	0.58	0.58	0.54	0.54
Support for Social-Emotional Development	0.61	0.56	0.55	0.65
Support for Language and Literacy Development	0.54	0.52	0.49	0.53
Support for Cognitive Development	0.61	0.64	0.55	0.47
Areas of Concern	0.54	0.57	0.75	0.65
Extreme Concern	0.69	--	--	0.61

Source: Q-CCIIT Fall 2012 Psychometric Field Test.

Note: The incidence of Extreme Concern was too low in centers to estimate correlations.

Summary. Overall, the Q-CCIIT measure demonstrated acceptable to strong inter-rater reliability, with slightly lower estimates for FCCs. Average item-level agreement across the entire rating form was 0.87, ranging from 0.84 for FCCs to 0.89 for center-based infant classrooms. On scale scores, agreement ranged from 0.87 to 0.90 for the full sample; the range for FCCs was lower (0.73 to 0.91), especially on Support for Language and Literacy Development. The average differences between observers on scales were usually half a point or less, with greater differences for FCCs. Weighted Kappas were acceptable for the full sample (0.54 to 0.71), although lower for Support for Language and Literacy Development with toddlers (0.49) and Support for Cognitive Development with FCCs (0.47). These findings may imply the need to provide more training on how to score when averaging interactions across children of different ages.

H. Reliability estimates from the perspective of generalizability theory

We use generalizability theory (G theory) (Brennan 2001a,b; Shavelson and Webb 1991) to examine simultaneously the different sources of error in measurement and estimate the number of cycles or observers needed to attain acceptable reliability using a dependability study (D-study, also known as a decision study).

In G theory, "a behavioral measurement score may be conceived of as a sample from a *universe of admissible observations*." A measurement situation has characteristic features, referred to as facets of measurement. G theory uses analysis of variance (ANOVA) procedures to obtain estimates of variance components for the different facets included in the analysis.

In the Q-CCIIT G-study, the classroom is the object of measurement, and the facets include observer, item, and cycle. Our data collection did not allow for the fully crossed models (cross-classified design with every observer observing the same classroom and occasions/cycles) used in many G-studies. For practical reasons, we designed the study with classrooms nested in

observers (that is, these variables were not completely crossed). Within each classroom, an observer rated the caregiver interactions multiple times (cycles) for most items on the Q-CCIIT instrument. The Q-CCIIT model assumes that the quality rating for a classroom is the true-score quality of the classroom, plus the effect of the observer, plus the interaction of the observer and classroom, in addition to the item interacting with occasion/cycle variance (the variance associated with the sampling of the time of observation).

For each of the three positive Q-CCIIT scales, our G-study analyses examined the variance component estimates for classrooms, observers, items, and cycles/occasions, and decomposed the estimate of variance components. The G-study (and D-study) analyses used standard ANOVA methods, in which the variance components were derived under the assumption of a balanced design (equal sample size for each facet). Because the number of Q-CCIIT classroom observations was not equal across observers (unbalanced nested design) and some data were missing (items not coded for particular cycles or activities), these results were exploratory rather than conclusive. Due to the nesting of classrooms within observers, the parameters could not be identified uniquely, so we needed to make an identifying assumption (Raudenbush et al. 2008). Because we expected that the observer effects would be stronger than the observer-by-classroom interactions, we assumed that the observer-by-classroom interaction would be 0. We investigated four different computational approaches for analyzing the Q-CCIIT data (Appendix F) to determine if there were any significant differences in results, but found only small differences between approaches; here we report the analysis using the Restricted Maximum Likelihood (REML) estimation method,[52] which allows an unbalanced design for estimating effects from the general linear mixed model (Patterson and Thompson 1971).

We conducted a series of multiple-facet, partially nested G-studies for each of the three Q-CCIIT scales identified by the factor analyses (Table VII.9). For each scale, we calculated variance components and their proportions of contribution for seven different models. Models 1 through 4 did not include the item term. Models 5 through 7 added the item term to examine the variance proportion contribution due to items and their interaction with the time sampling (cycles). The model specifications follow:

Model 1: Classrooms nested within observers; observers; and the residual term.

Model 2: Classrooms nested within observers; cycles; observers; and the residual term.

Model 3: Classrooms nested within observers; cycles; observers; cycle by observer interaction; and the residual term.

Model 4: Classrooms nested within observers; cycles; observers; cycle by observer interaction; cycle by classroom nested within observer interaction; and the residual term.

Model 5: Classrooms nested within observers; items; cycles; observers; and the residual term.

[52] For the REML method, we specified the maximum number of iterations to be 100.

Model 6: Classrooms nested within observers; items; cycles nested within observers; observers; and the residual term.

Model 7: Classrooms nested within observers; items; cycles; observers; item by cycle interaction; item by observer interaction; cycle by observer interaction; item by classroom nested within observer interaction; cycle by classroom nested within observer interaction; and the residual term.

Model 7 results are described in Table VII.7.

Table VII.7. Q-CCIIT G-study results for positive scales

	Social-emotional development	Language and literacy development	Cognitive development
Source of variation	Percentage contribution	Percentage contribution	Percentage contribution
Classroom (Observer)	38	25	25
Observer	6	11	9
Item	3	13	4
Cycle	<0.01	<0.01	0
Observer x Item	4	5	6
Observer x Cycle	<0.01	<1	1
Item x Cycle	<0.01	<1	<0.01
Item x Class (Observer)	12	7	10
Cycle x Class (Observer)	19	18	13
Residual	17	22	32

Source: Q-CCIIT Fall 2012 Psychometric Field Test.

Support for Social-Emotional Development. The greatest proportion of the variance for this skill was accounted for by the classroom nested within observer term, which is the object of measurement. Here, 69 percent of the variance for the social-emotional development skill could be accounted for by individual classrooms and the interactions of classrooms with items and cycles (with an additional 3 percent attributable to the items), taking into account that observers were assigned to specific groups of classrooms for observations. The variance attributed to observers and the interaction of observers with items and cycles was 10 percent. In addition, 17 percent of the variance could not be accounted for in this model (residual). These results indicate that classroom caregivers (nested within observer) varied in their ability to provide support for the social-emotional development of the children in their classrooms. The classroom level of quality, and not the rating style of the observer, accounted for most of the variance.

Support for Language and Literacy Development. Though not as strong as the Support for Social-Emotional Development scale, results for the Support for Language and Literacy Development scale also indicate that classrooms nested within observers accounted for the greatest proportion of the variance. Fifty percent of the variance was accounted for by individual classrooms and the interactions of classrooms with items and cycles (with an additional 13 percent attributable to the items), taking into account that observers were assigned to specific

groups of classrooms for observations. The variance attributed to observers and the interaction of observers with items and cycles was 16 percent. A greater portion of the variance than found in the previous scale (residual of 22 percent) could not be accounted for by individual classrooms, observers, items and cycles measurements, and their two-way interaction terms in Model 7. The classroom level of quality, and not the rating style of the observer, accounted for most of the variance.

Support for Cognitive Development. Although the scale for Support for Cognitive Development had the weakest evidence of generalizability, with a substantial portion of the variance (residual of 32 percent) not accounted for by the model, the results indicate that classrooms nested within observers varied in caregivers' ability to provide support for the cognitive development of the children in these classrooms, and that it was the classroom level of quality, rather than the rating style of the observer, that accounted for most of the variance. The proportion of the variance on the Support for Cognitive Development scale accounted for by the classrooms nested within observer and the classroom interactions with items and cycles was 48 percent, with an additional 4 percent attributable to the items. The variance attributed to observers and the interaction of observers with items and cycles was 16 percent.

Patterns of Unaccounted-For Variance. The unaccounted-for variance in the models could be related to the day of the week, the content of activities, or the differences in classroom composition. The findings from descriptive analyses of differences in caregiver interactions with infants versus toddlers suggest that some of the variance in Q-CCIIT may be related to the ages of the children involved in interaction with the caregivers. In addition, our qualitative observations during the pretesting of Q-CCIIT suggested that more verbal and assertive children received more attention, and often greater support for language and cognition.

D-Study. In the Q-CCIIT design, it is difficult to conduct a D-study due to the use of a nested design, particularly when classrooms (the measurement object) were nested within observers (facet). We did want to explore how much we could increase reliability of our classroom estimates by adding cycles of observation within a visit. For each of the Q-CCIIT scales, we calculated the statistics for our D- studies based on an assumption that the sample sizes are equal/balanced; that is, within each classroom, n'_o number of observers observed n'_y cycles, and that the items used are those defined for each of the three Q-CCIIT scales. Under this assumption, the values for the estimates of the G-coefficient are provided for each of the three Q-CCIIT scales (Figures VII.3 through VII.5).

Under the current design (and using results from our G-study, assuming that the number of items used is the same as in the current study, and at least five cycles were observed by an observer within a classroom), the G-coefficient and dependability index (phi) for each of the three support skill areas show a good level of reliability; that is, they are mostly greater than 0.55. For example, the G-coefficient for Support for Social-Emotional Development, Support for Language and Literacy Development, and Support for Cognitive Development when one observer was used to observe five cycles were, respectively, 0.87, 0.84, and 0.84. With only four cycles, the G-coefficients would be 0.85, 0.81, and 0.81 respectively.

The results of the dependability study suggest that the phi (dependability index) for the Support for Social-Emotional Development, Support for Language and Literacy Development, and Support for Cognitive Development with five cycles and one observer would be 0.75, 0.59,

and 0.63, respectively, while the same number of cycles with two observers would be 0.85, 0.73, and 0.77, respectively.

Figure VII.3. G-coefficient: D-study results under model 7 for Q-CCIIT support for social-emotional development

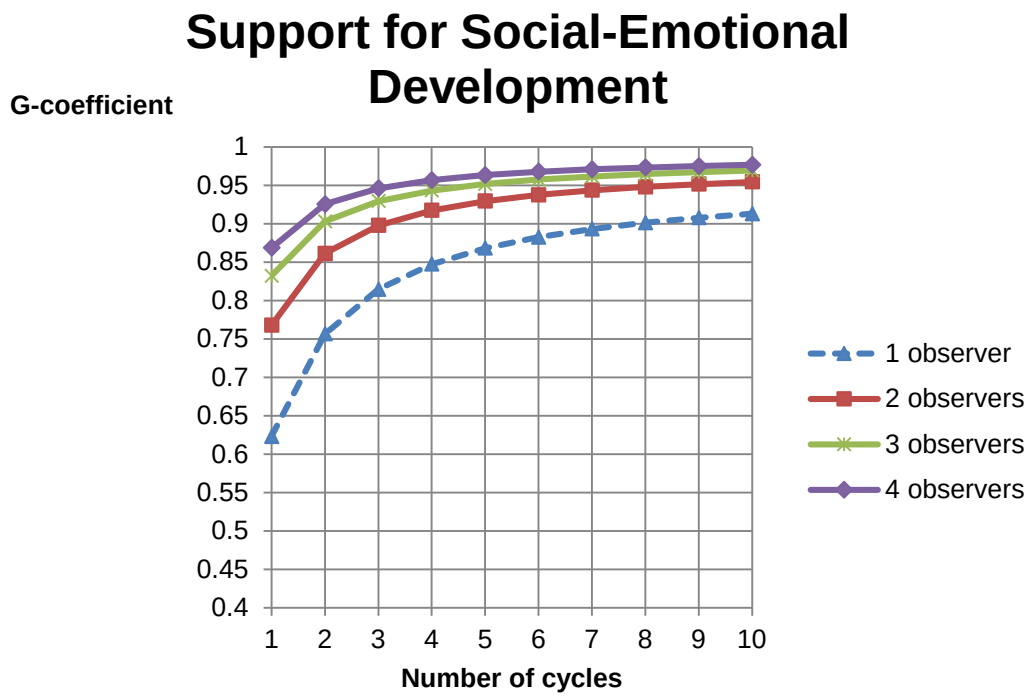


Figure VII.4. G-coefficient: D-study results under model 7 for Q-CCIIT support for language and literacy development

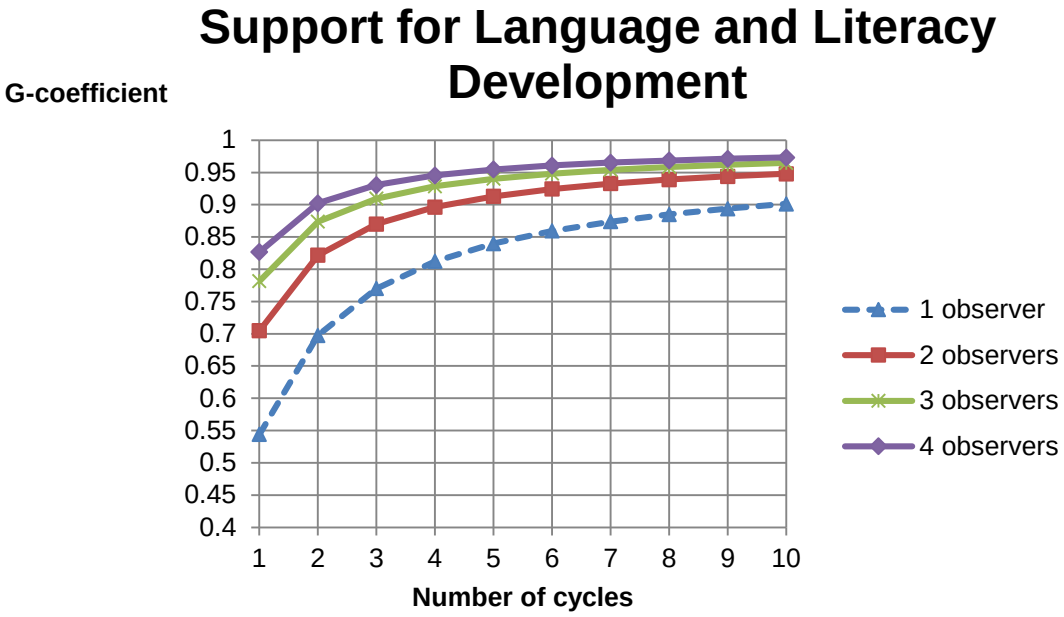
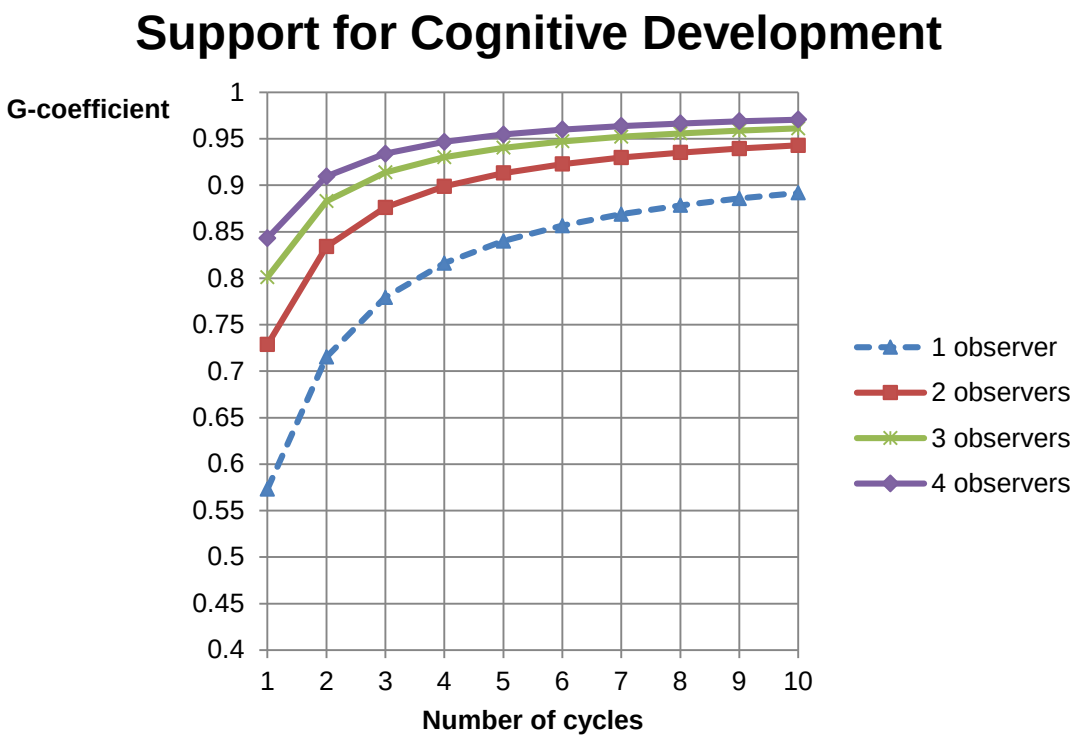


Figure VII.5. G-coefficient: D-study results under model 7 for Q-CCIIT support for cognitive development



Summary. The G-study indicated that most of the variance may be attributed to differences in classrooms (25 to 38 percent, and the interaction of classrooms with items and cycles (48 to 69 percent). Observer-alone and observer by item and cycle variance were relatively low. Residual variance ranged from 17 to 32 percent; thus, this model accounted for more of the variance found in the measures than other studies of classroom observations. For example, G-studies of observations in elementary schools, such as the Literacy Observation Tool Reliability Study (Grehan and Sterbinsky 2005) and Measuring Effective Teaching study (MET study; Kane and Staiger 2005) , have had residual variance ranging from 32 to 82 percent. D-study results indicated that the G-coefficient and dependability index (phi) for each of the three support skill areas showed a good level of reliability, with most greater than 0.55.

I. Confirmatory Factor Analysis (CFA)

The results of the CFA provided evidence for the construct validity of the Q-CCIIT scales. With CFA, we tested hypotheses corresponding to the theoretical notions of the quality of caregiver-child interactions for infants and toddlers using the field test data; we used Mplus (Muthén and Muthén 2007) to examine the associations between observed indicators and primary latent factors, and the correlations between latent factors for the full sample. The model fit estimates assess how well the structure of scales on the new quality measure and the items in each subscale capture the covariance between all items on the measure. Poor fit, if revealed, may be attributable to the presence of some items measuring multiple factors or some items within a factor being more related to each other than to others (Brown 2006).

By convention, rules of thumb for a well-fitting model are comparative fit index (CFI) and Tucker-Lewis index (TLI) greater than 0.90 and root mean square error of approximation (RMSEA) and standardized root mean square residuals (SRMRs) less than 0.10. We use these criteria to determine whether the model is acceptable.

For each item, we averaged across the observation cycles and used the item averages as indicators in the model. Exceptions were those items rated at the end of the observation (or book-sharing items that may occur in a single cycle), for which we used the single item rating.

In addition to performing CFA for the full sample of classrooms and providers, we also examined factor structure by program type (centers and FCCs), child age (infant and toddler), and concentration of DLLs in the classrooms (high and low concentration)[53] to test whether the factor structure of the new measure is the same in specified subgroups. The sample sizes for the subgroup analyses fall below the recommended size for CFA, so the findings for the subgroups are promising but not definitive. We noted differences in loading > 0.10 between subgroups but did not test for differences, given the sample size.

CFA with the Full Sample. Fit statistics for the CFA based on the initially proposed scales (Table V.2) indicated poor model fit; for this reason, we conducted exploratory factor analysis (EFA) as a preliminary step to inform the CFA. The EFA results indicated the presence of a method effect. Within each domain, the items collected repeatedly (in each of the observation cycles) and the items rated across the visit (a single rating) tended to load on separate factors. In addition, three items loaded on scales not proposed in the analysis plan. The item about supervising and joining children's play, initially proposed as supporting cognitive development, loaded on the Support for Social-Emotional Development scale. The two items assessing peer interactions initially proposed for the Supporting for Social-Emotional Development scale loaded on the Support for Cognitive Development scale. The peer interactions items involve helping children understand that peers are people rather than toys and learn how to interact and solve problems with peers—which given the results of the loadings indicate these are more conceptual items, not relational. Caregivers who supported problem solving and other types of play were more likely to support peer interactions than those who were not as supportive of cognitive development. By including the two peer interaction items in the Support for Cognitive Development scale, the Support for Social-Emotional Development scale is now a measure of the responsiveness and relationship-building interactions of the adult caregivers in a classroom with the children (rather than facilitation of peer social skills). On the other hand, the Support for Cognitive Development scale examines introduction of new concepts and ideas, and support for different types of play, with an emphasis on exploration and problem solving, including peer interactions. The individual Types of Talk variables did not load on the Support for Language and Literacy Development. With very limited variance, the Areas of Concern items did not form a factor.

Based on these EFA results, we re-estimated the CFA, reassigning the indicator items for the social-emotional and cognitive scales and separating them by the across-the-visit factor and observation cycle factors into subscales of the three developmental support scales. The fit statistics were adequate for the full sample (Table VII.8: CFI = 0.94; TLI = 0.93; RMSEA =

[53] High concentration of DLLs is defined as 50 percent or more children with non-English home language, as reported by caregivers in the Classroom Roster.

0.06; SRMR = 0.04), suggesting that the model fit the data well. Table VII.9 identifies the items included in each of the scales based on the CFA.

Table VII.8. Model fit statistics for the full sample, and by subgroups

		Program type		Child age		DLL concentration	
	Full sample	FCC	Center	Infant	Toddler	High	Low
CFI	0.94	0.93	0.94	0.93	0.92	0.92	0.93
TLI	0.93	0.92	0.93	0.92	0.91	0.91	0.92
RMSEA	0.06	0.07	0.06	0.07	0.07	0.08	0.07
SRMR	0.04	0.06	0.05	0.06	0.06	0.06	0.05

Source: Q-CCIIT Fall 2012 Psychometric Field Test.

Table VII.9. Latent factors/scales based on CFA and observed indicators

Support for social emotional development	Support for cognitive development	Support for language and literacy development
Responding contingently to distress	Supporting object exploration	Use of varied vocabulary
Responding contingently to social cues	Scaffolding problem solving	Use of questions
Responding to emotional cues	Extending pretend play[a]	Conversational turn taking
Building a positive relationship	Explicit teaching[a]	Extending children's language use
Responsive routines[a]	Giving choices[a]	Engaging children in books
Classroom limits and management [a]	Support for social problem solving[a]	Variety of words (book sharing)
Sense of belonging[a]	Supporting peer interaction/play	Variety of types of sentences (book sharing)
Supervises or joins in play and activities[a]	Diversity of concepts	Features of talk[a]
		Talk about things not present[a]
		Positive attitudes toward books[a]

Source: Q-CCIIT Fall 2012 Psychometric Field Test.

Note: Indicators are sometimes based on multiple items.

[a] Items rated at the end of the observation.

Figure VII.6 shows the final factor structure and observed indicator items for the full sample. There were two factors for the Support for Social-Emotional Development scale: the observation cycle factor included responding contingently to distress, responding contingently to social cues, responding to emotional cues, and building a positive relationship (factor loadings ranged from 0.69 to 0.95); the across-the-visit factor included responsive routines, classroom limits and management, sense of belonging, and supervises or joins in play and activities (factor loadings ranged from 0.76 to 0.83). There were three factors for the Support for Language and Literacy Development scale: the observation cycle factor included use of varied vocabulary, conversational turn taking, use of questions, and extending children's language use (factor

loadings ranged from .84 to .90); the across-the-visit factor included features of talk, talk about things not present, and positive attitudes toward books (factor loadings ranged from 0.63 to 0.77); the book-sharing factor included engaging children in books, variety of words, and variety of types of sentences (factor loadings ranged from 0.76 to 0.89). Only one factor was formed for Support for Cognitive Development; it included diversity of concepts (number of unique concepts across cycles), extending pretend play, scaffolding problem solving, explicit teaching, giving choices, supporting peer interaction/play, support for social problem solving, and supporting object exploration (factor loadings ranged from 0.54 to 0.80).

The moderate to high inter-factor correlations suggest that second order factors might exist. The inter-factor correlations ranged from 0.56 to 0.91, with most of the correlations greater than 0.70 (Table VII.10). We estimated a second-order factor analysis by grouping the two social-emotional factors for Support for Social-Emotional Development and the three language factors for Support for Language and Literacy Development. The fit statistics indicated good model fit (CFI = 0.94; TLI = 0.93; RMSEA = 0.06; SRMR = 0.05). These results suggest that there are three overall Q-CCIIT scales, as we proposed, although the items in each scale are slightly different from what we proposed. The factor loadings were 0.84 and 0.92 for the observation cycle factor and the across-the-visit factor, respectively, for Support for Social-Emotional Development. The factor loadings ranged from 0.67 to 0.93 for the three language factors on the second order Support for Language and Literacy Development factor.

Inter-factor correlations suggest that a total score could be estimated. The total Q-CCIIT score would represent supportive, responsive caregiving across different areas of support for development.

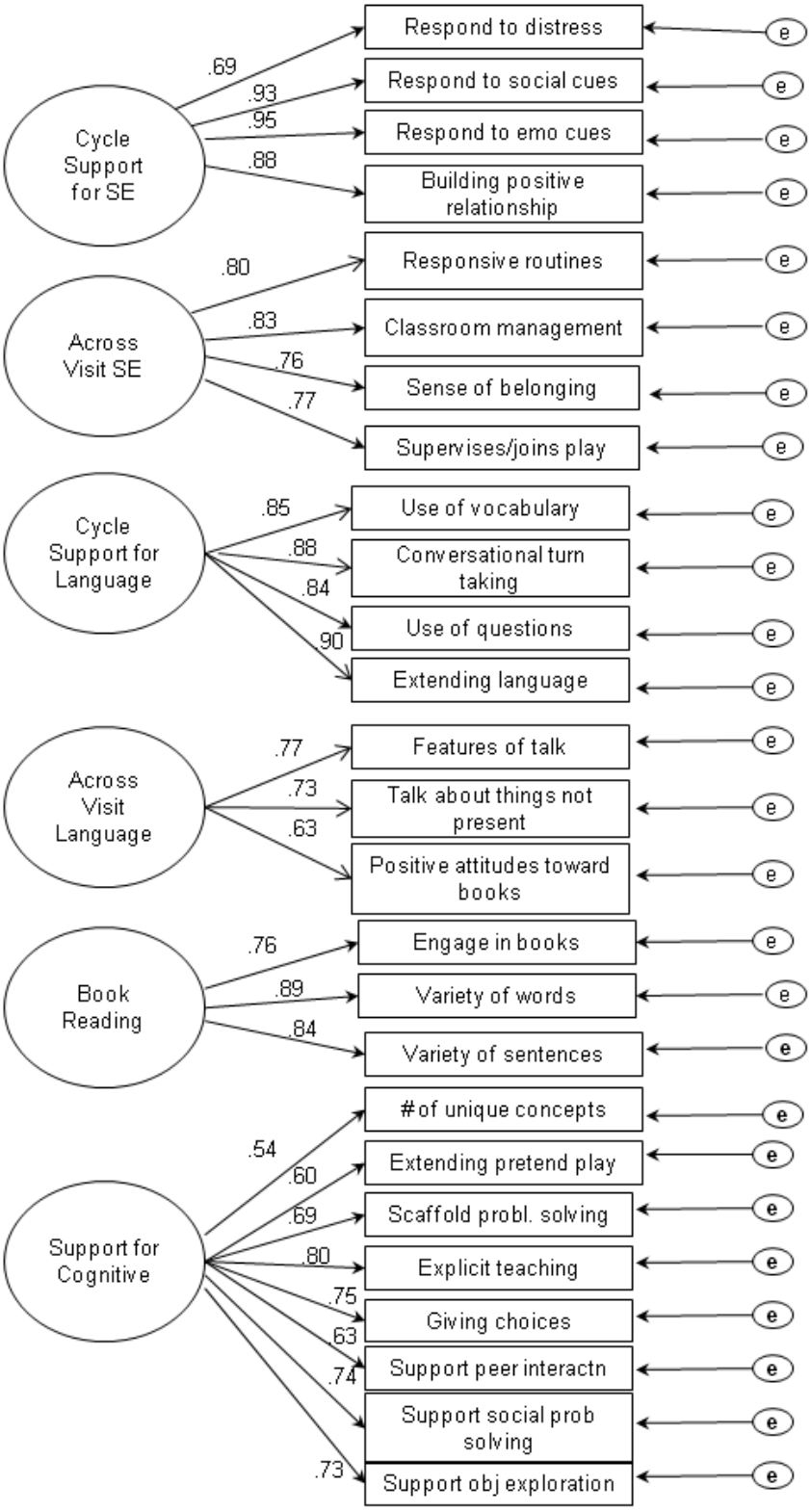
Table VII.10. Inter-factor correlations for the full sample

	Cycle support for social-emotional	Across-the-visit social-emotional	Cycle support for language	Across-the-visit language	Book sharing
Across-the-Visit Social-Emotional	.78				
Cycle Support For Language	.74	.77			
Across-the-Visit Language	.73	.85	.83		
Book Sharing	.59	.61	.63	.67	
Cycle Support For Cognitive	.74	.84	.91	.85	.56

Source: Q-CCIIT Fall 2012 Psychometric Field Test.

CFA by Subgroups. We ran the CFA by program type, child age, and concentration of DLLs. Factor structure was the same in the FCCs and centers (Figure VII.7), in the infant and toddler classrooms (Figure VII.8), and in the high and low DLL classrooms (Figure VII.9). Model fit statistics were similar in FCCs and centers, infant and toddler classrooms, and high and low concentration of DLL classrooms (Table VII.8).

Figure VII.6. Confirmatory factor analysis: full sample



Inter-factor correlations range from 0.56 to 0.91

Figure VII.7. Confirmatory factor analysis: centers and FCCs

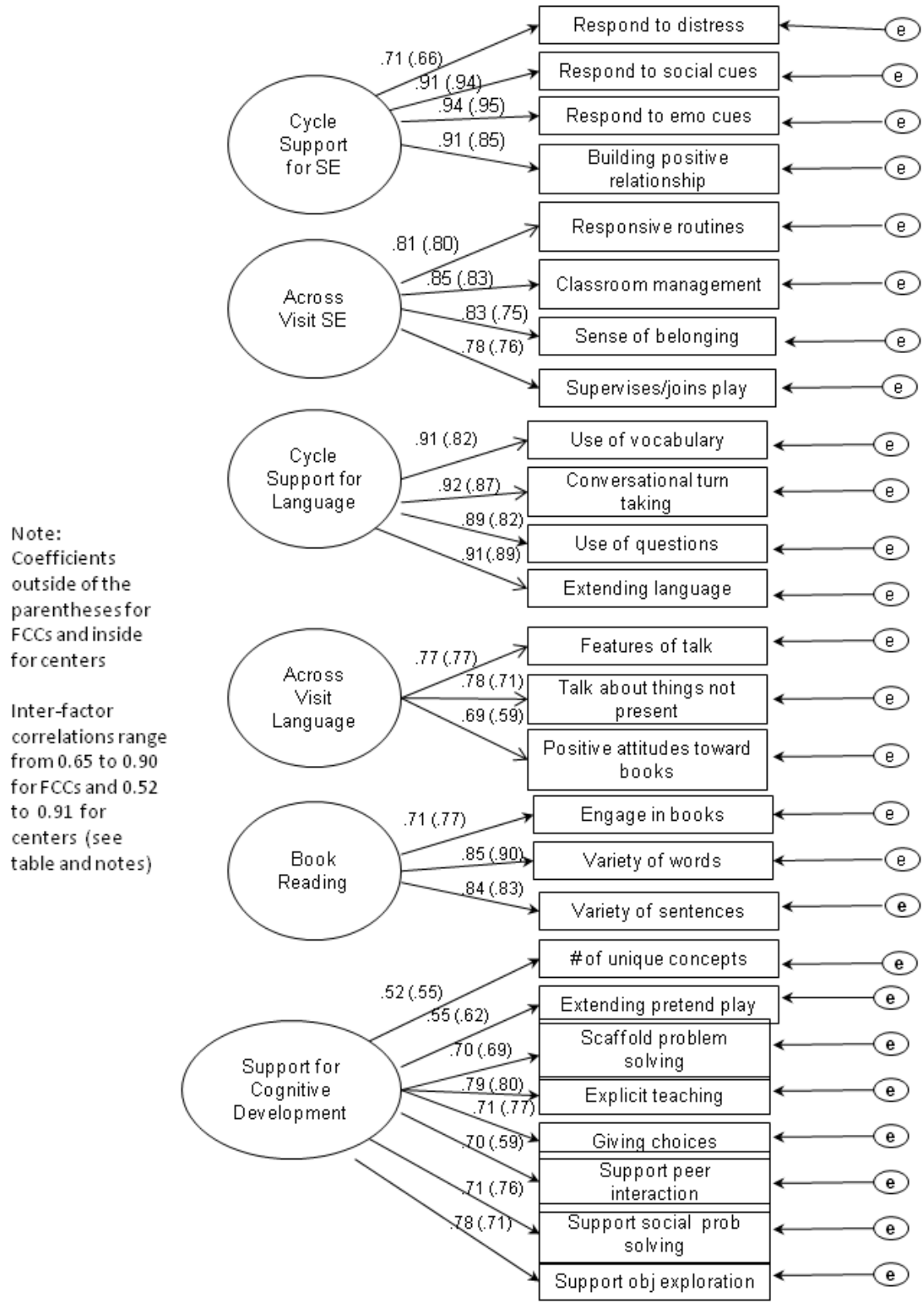
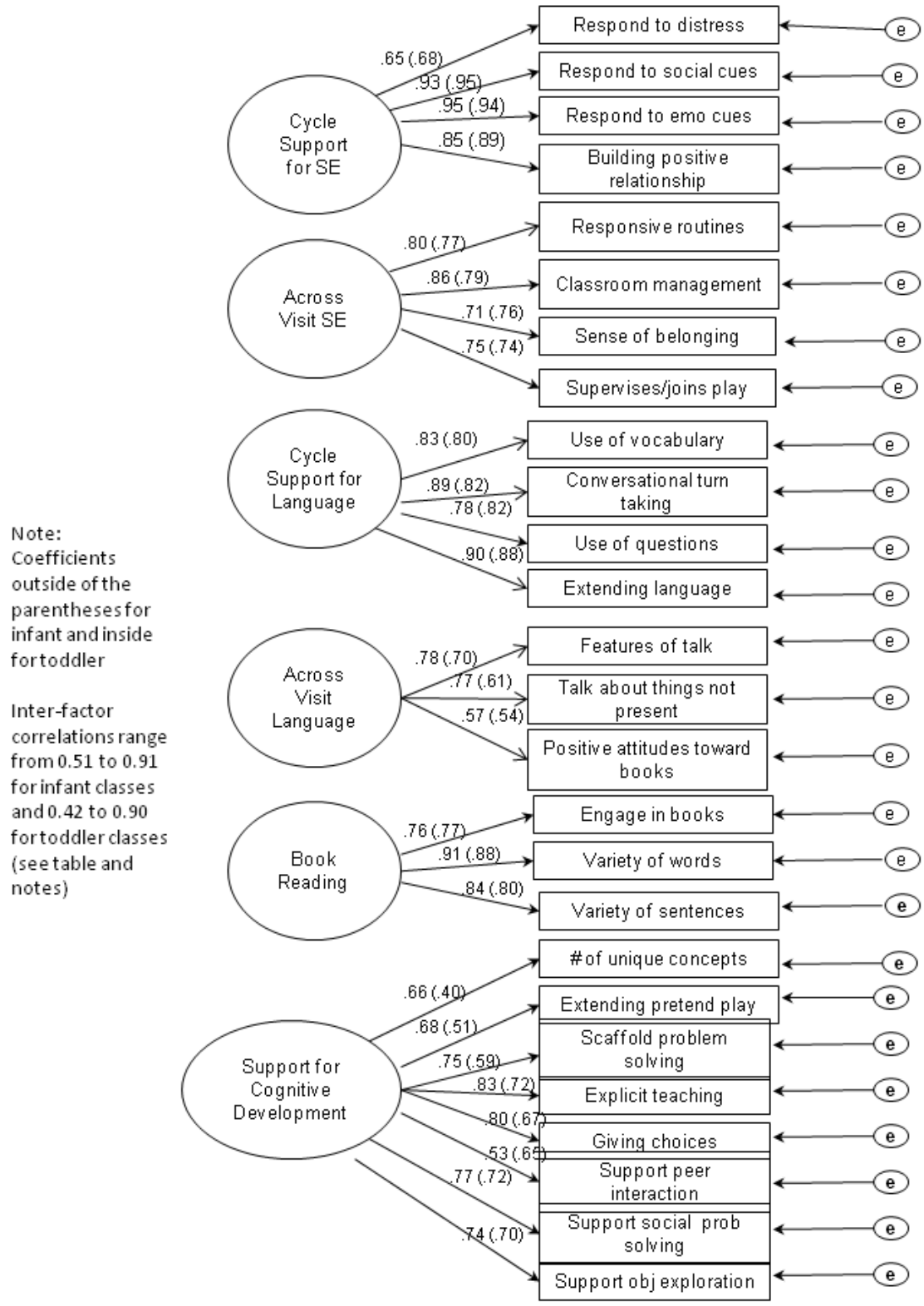


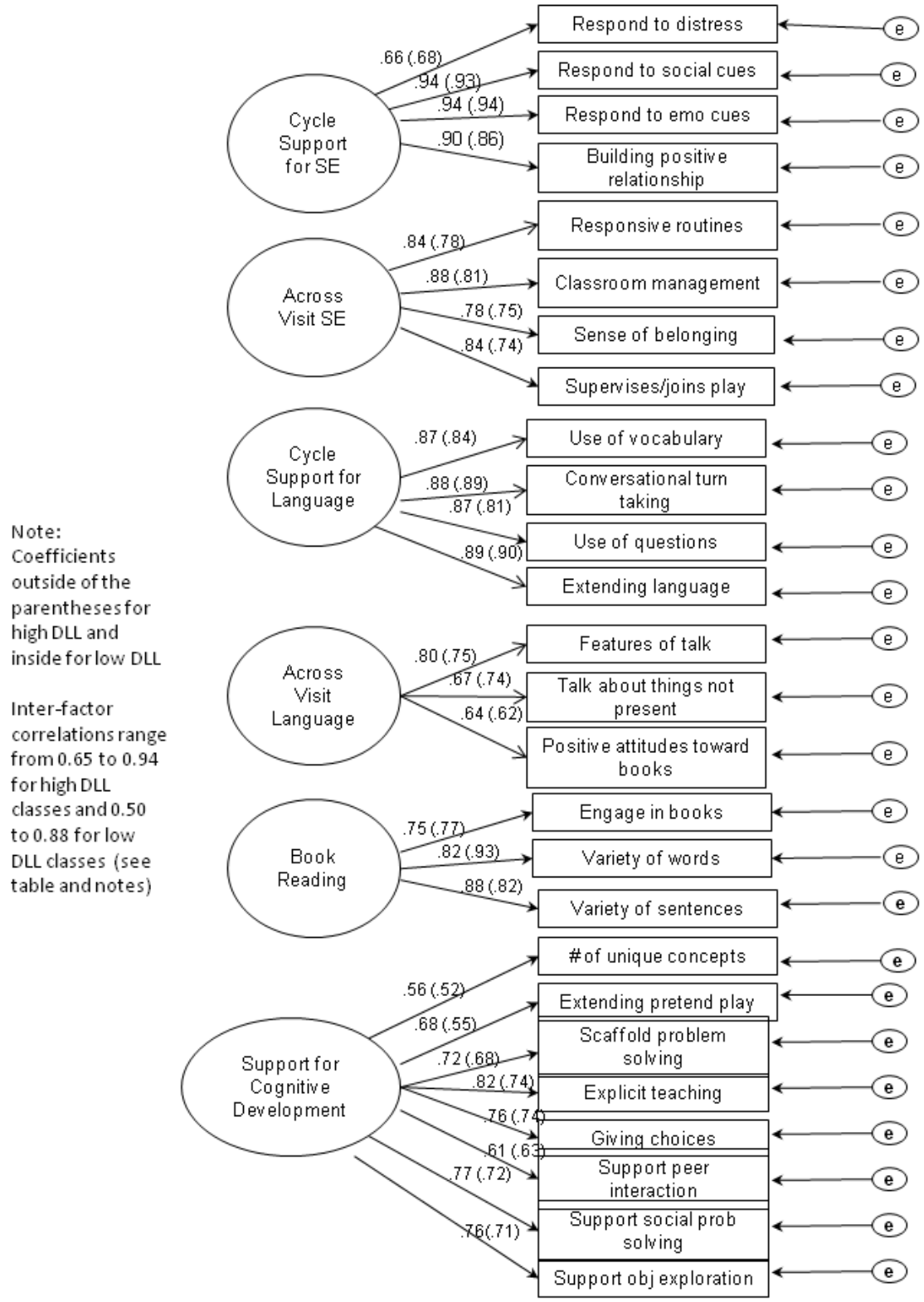
Figure VII.8. Confirmatory factor analysis: infant and toddler classrooms



Note:
Coefficients outside of the parentheses for infant and inside for toddler

Inter-factor correlations range from 0.51 to 0.91 for infant classes and 0.42 to 0.90 for toddler classes (see table and notes)

Figure VII.9. Confirmatory factor analysis: high and low concentration DLL classrooms



Note:
Coefficients outside of the parentheses for high DLL and inside for low DLL

Inter-factor correlations range from 0.65 to 0.94 for high DLL classes and 0.50 to 0.88 for low DLL classes (see table and notes)

FCCs vs. Centers. Generally, the factor loadings were similar in FCCs and centers (with differences in factor loadings < 0.10), with two exceptions: the factor loading was higher in FCCs than in centers for positive attitudes toward books (0.69 versus 0.59) and supporting peer interaction/play (0.70 versus 0.59) (Figure VII.7).

Table VII.11 presents the inter-factor correlations for FCCs and centers. (Correlations for FCCs are listed first; centers are listed in parentheses. Bold numbers indicate that the difference in *r* was greater than 0.10.) Book Sharing was more strongly correlated with the across-the-visit language factor for FCCs ($r = 0.75$) than centers ($r = 0.64$); it was also more strongly correlated with the cognitive factor for FCCs ($r = 0.67$) than for centers ($r = 0.52$). Other correlations were similar between FCCs and centers.

Table VII.11. Inter-factor correlations for FCCs and centers

	Cycle support for social-emotional	Across-the-visit social-emotional	Cycle support for language	Across-the-visit language	Book sharing
Across-the-Visit Social-Emotional	.80 (.76)	--			
Cycle Support for Language	.79 (.73)	.79 (.78)	--		
Across-the-Visit Language	.67 (.76)	.84 (.86)	.78 (.86)	--	
Book Sharing	.65 (.59)	.66 (.61)	.70 (.61)	**.75 (.64)**	--
Cycle Support for Cognitive	.80 (.72)	.89 (.84)	.90 (.91)	.82 (.87)	**.67 (.52)**

Source: Q-CCIIT Fall 2012 Psychometric Field Test.

Note: Correlations for FCCs are listed first; centers are listed second and in parentheses. Bold numbers indicate that the difference in *r* was greater than 0.10.

Infant vs. toddler classrooms. More items differed in factor loadings between infant and toddler classrooms than in the FCC vs. center comparisons, with the loadings usually stronger in infant than toddler classrooms: for example, one item for the Across-the-Visit Language factor—talk about things not present (0.77 versus 0.61)—and four items for the Cycle Support for Cognitive Development factor—number of unique concepts (0.66 versus 0.40), extending pretend play (0.68 versus 0.51), explicit teaching (0.83 versus 0.72), and giving choices (0.80 versus 0.67). One exception was support for peer interaction/play, for which the loading was higher in toddler than infant classrooms (0.65 versus 0.53). The factor loadings were similar in infant and toddler classrooms for the remaining items (Figure VII.8).

Table VII.12 presents the inter-factor correlations for infant and toddler classrooms. (Correlations for toddler classrooms are listed first; infant classrooms are listed second and in parentheses. Bold numbers indicate that the difference in *r* was greater than 0.10.) Book Sharing was more strongly related to Cycle Support for Social Emotional Development for toddlers ($r = 0.65$) than infants ($r = 0.54$); Cycle Support for Language was more strongly related to the Cycle Support for Social-Emotional Development for toddlers ($r = 0.79$) than infants ($r = 0.69$). Other correlations were similar between infant and toddler classrooms.

Table VII.12. Inter-factor correlations for infant and toddler classrooms

	Cycle support for social-emotional	Across-the-visit social-emotional	Cycle support for language	Across-the-visit language	Book sharing
Across-the-Visit Social-Emotional	.78 (.76)	--			
Cycle Support for Language	**.79 (.69)**	.78 (.76)	--		
Across-the-Visit Language	.80 (.77)	.88 (.84)	.88 (.82)	--	
Book Sharing	**.65 (.54)**	.56 (.61)	.59 (.56)	.60 (.60)	--
Cycle Support for Cognitive	.76 (.74)	.84 (.83)	.90 (.91)	.86 (.83)	.42 (.51)

Source: Q-CCIIT Fall 2012 Psychometric Field Test.

Note: Correlations for toddlers are listed first; infants are listed second and in parentheses. Bold numbers indicate that the difference in *r* was greater than 0.10.

High vs. Low Concentrations of DLLs. The factor loadings between classrooms with high and low concentrations of DLLs were similar (Figure VII.9), with two exceptions: the factor loading for extending pretend play was stronger in high than low DLL classrooms (0.68 versus 0.55), and the factor loading for variety of words was weaker in high than low DLL classrooms (0.82 versus 0.93).

Table VII.13 presents the inter-factor correlations for high and low DLL classrooms. (Correlations for low DLL classrooms are listed first; high DLL classrooms are listed second and in parentheses. Bold numbers indicate that the difference in *r* was greater than 0.10.) The correlations generally were stronger in high (0.65 to 0.90) than low DLL classrooms (0.57 to 0.88). For example, Across-the-Visit Language was more strongly related to the Cycle Support for Social-Emotional Development (0.90 versus 0.65), Cycle Support for Cognitive Development (0.94 versus 0.81), and Book Sharing (0.81 versus 0.60) in high than low DLL classrooms. Similarly, Book Sharing and Support for Cognitive Development were more strongly related for high DLL (0.67) than low DLL (0.50) classrooms. Support for Cognitive Development and Cycle Support for Social-Emotional Development also had a stronger correlation in high (0.83) than low DLL (0.69) classrooms.

Table VII.13. Inter-factor correlations for high and low DLL classrooms

	Cycle support for social-emotional	Across-the-visit social-emotional	Cycle support for language	Across-the-visit language	Book sharing
Across-the-Visit Social-Emotional	.77 (.80)	---			
Cycle Support for Language	**.**73 (.77)	**.**74 (.77)	--		
Across-the Visit-Language	**.65 (.90)**	.82 (.90)	**.79 (.90)**	--	
Book Sharing	**.**57 (.65)	.59 (.67)	**.60 (.72)**	**.60 (.81)**	--
Cycle Support for Cognitive	**.69 (.83)**	.84 (.85)	**.**88 (.94)	**.81 (.94)**	**.50 (.67)**

Source: Q-CCIIT Fall 2012 Psychometric Field Test.

Note: Correlations for low DLL classrooms are listed first; high DLL classrooms are listed second and in parentheses. Bold numbers indicate that the difference in *r* was greater than 0.10.

Summary. CFA results indicated similar factor structures by subgroups. Fit statistics for the original CFA, based on the initially proposed scales, indicated poor model fit; for this reason, we conducted an EFA step to inform the CFA. The EFA results indicated the presence of a method effect. That is, within each domain, the items rated across the observation cycles and those rated across the visit tended to load on separate factors. In addition, three items loaded on scales differently than we originally proposed. The fit statistics for the new CFA indicated adequate model fit for the full sample. The factors were moderately to highly correlated with each other; most of the correlations exceeded 0.70, suggesting that second-order factors may exist. When estimating a second-order factor analysis (by grouping the two social-emotional factors for Support for Social-Emotional Development and the three language factors for Support for Language and Literacy Development), the fit statistics indicated good model fit. These results suggest that there are three overall Q-CCIIT scales, as proposed, although the items in each scale are slightly different from those originally proposed. In addition, the high inter-factor correlations help support the possibility of estimating a total Q-CCIIT score, which would represent supportive, responsive caregiving across different areas of support for development.

Regarding CFA for subgroups, model fit statistics were similar in FCCs and centers, infant and toddler classrooms, and high and low concentration DLL classrooms, indicating good model fit. Generally, the factor loadings were similar in FCCs and centers. More items differed in factor loadings between infant and toddler classrooms, with the loadings stronger in infant than toddler classrooms for five items and weaker in one item. The factor loadings between classrooms with high and low concentrations of DLLs were similar, with the exception of two items. Thus, CFA results indicate similar factor structures by subgroups.

J. Item-Response Theory (IRT) analysis

We used item-response theory in analyzing the field test data to assess reliability and construct validity of the measure for three of the Q-CCIIT scales[54] based on the CFA results (Support for Social-Emotional Development, Support for Cognitive Development, and Support for Language and Literacy Development). We applied a one-parameter Rasch rating scale model to each scale.[55] In the Rasch model, the probability of a specified response is modeled as a function of a classroom quality level for the measured construct and item difficulty. Item difficulty[56] and classroom quality are placed on the same scale and expressed as log odds, providing an interval measurement scale.

[54] The Areas of Concern scale had very limited variance on most items (and most of that is from FCCs), so we did not attempt an IRT analysis.

[55] Rasch models assume unidimensionality. Responsive caregiving is the underlying dimension of all of the scales, and we analyzed the scales individually (by the domain supported). Results indicated that we met the unidimensionality criteria, although these analyses also identified a method difference in the factor analyses of the residuals.

[56] In our analysis of the Q-CCIIT measure, item difficulty refers to the difficulty in reaching higher levels of quality on the item. The quality practices that are easier to implement (and thus more prevalent across settings) are considered low difficulty items; those quality practices more difficult to implement are considered higher difficulty items.

A sample size of 30 items and 30 classrooms is usually adequate for estimating the difficulty of items and the quality of classrooms within one logit, with 95 percent confidence (Linacre 1994). At least 10 observations are recommended for each category per item to estimate the step parameters (the distance between categories) with confidence. The field test sample (N = 400) and each of the subsamples (N > 100, range = 110 to 290) were large enough to provide reliable estimates of the step parameters and item difficulties.

Reliability estimates are provided for both item difficulty and classroom quality estimates. For item difficulty, the IRT reliability estimates ranged from 0.97 to 0.99 overall and by subgroups, indicating excellent reliability. Table VII.14 presents the IRT reliability of the classroom quality estimates for the three Q-CCIIT scales for the full sample and by child age and program type. The reliability estimates ranged from 0.88 to 0.92 for the full sample, indicating strong reliability. The estimates were similar between infant and toddler classrooms, FCCs and centers, and high and low DLL classrooms. These are consistent with the other reliability estimates (that is, coefficient alpha and G-coefficients).

Table VII.14. IRT reliability estimates for the Q-CCIIT scales, for the overall sample and by child age and program type

		Child age		Program type		DLL concentration	
Q-CCIIT Scales	Full sample	Infant	Toddler	Center	FCC	High	Low
Support for Social-Emotional Development	.92	.92	.92	.92	.93	.93	.91
Support for Language and Literacy Development	.92	.91	.90	.91	.93	.93	.91
Support for Cognitive Development	.88	.89	.85	.88	.89	.90	.87
Sample Size	**400**	**136**	**154**	**290**	**110**	**126**	**253**

Source: Q-CCIIT Fall 2012 Psychometric Field Test.

Rasch models can provide evidence of the construct validity of the measure by illustrating whether the ordering of the items is consistent with theory about the construct. For example, caregiver interactions that theoretically require greater skill to implement regularly (such as scaffolding problem solving) should have greater item difficulty estimates than interactions considered easier to implement regularly (such as supporting object exploration). If the items measure a construct reliably, the hierarchy of items should be consistent with theoretical assumptions about the construct.

Fit statistics are generally provided for step increases (the distance between categories), items, and classrooms. They are used to identify strong items as well as any problems in measurement. Fit statistics signal items that are not contributing to measurement and can help to identify categories that are used interchangeably and thus do not indicate a real increase in quality.

The results from the Rasch analysis indicated that the ordering of item difficulties was consistent with theoretical assumptions about the relative likelihood of occurrence of high quality interactions—for example, quality support for object exploration was more likely to be observed than quality support for scaffolding problem solving. Overall, the item fit statistics were in the acceptable range, indicating good fit to the model, with the exception of extending representational play. Extension of representational play seldom was observed, so estimates of item difficulty are less reliable (that is, some caregivers may score well on representational play although they did not score well on less difficult skills, such as explicit teaching). However, given its role in early development and our plan to use mean scores rather than Rasch scores, we retained this item. The mean and Rasch scores were highly correlated; in order to test convergent and discriminant validity, we chose to use a score that most future users would be able to estimate.

Figures VII.10 to VII.12 display the item ordering for the three scales (on the right of the scale). From the bottom to the top in the figures, the items are ordered from the easiest to the most difficult. The easiest item suggests that caregivers had a high probability of receiving a high rating on this item; the most difficult item suggests that caregivers seldom received a high rating on this item (in other words, only caregivers rated highly on other items would be expected to receive a high rating on this item). The left side of the figure (a histogram) illustrates the distribution of scale scores. These results indicate that the field test sample included classrooms and FCCs with a range of quality, with Support for Social-Emotional Development being the most normally distributed. The items are matched to the overall quality of the items—particularly for Support for Social-Emotional Development and Support for Language and Literacy Development. The M on the left side of the dotted line indicates the mean for the classrooms, while the M on the right side indicates the mean difficulty of the items (with S = 1 standard deviation and T = 2 standard deviations).

For the Support for Social-Emotional Development scale (Figure VII.10), the easiest item is building a positive relationship; the most difficult item is classroom limits and management.

For the Support for Language and Literacy Development scale (Figure VII.11), the easiest item is engaging children in books; the most difficult item is talk about things not present. As expected when we were selecting the task of sharing a book with children, the language used by caregivers was more complex and varied during book sharing than at other times. Items addressing eliciting and extending child language are among the more difficult items.

For the Support for Cognitive Development scale (Figure VII.12), the easiest item is supporting object exploration; the most difficult item is extending pretend play.

Figure VII.10. Item map for Q-CCIIT support for social-emotional development scale

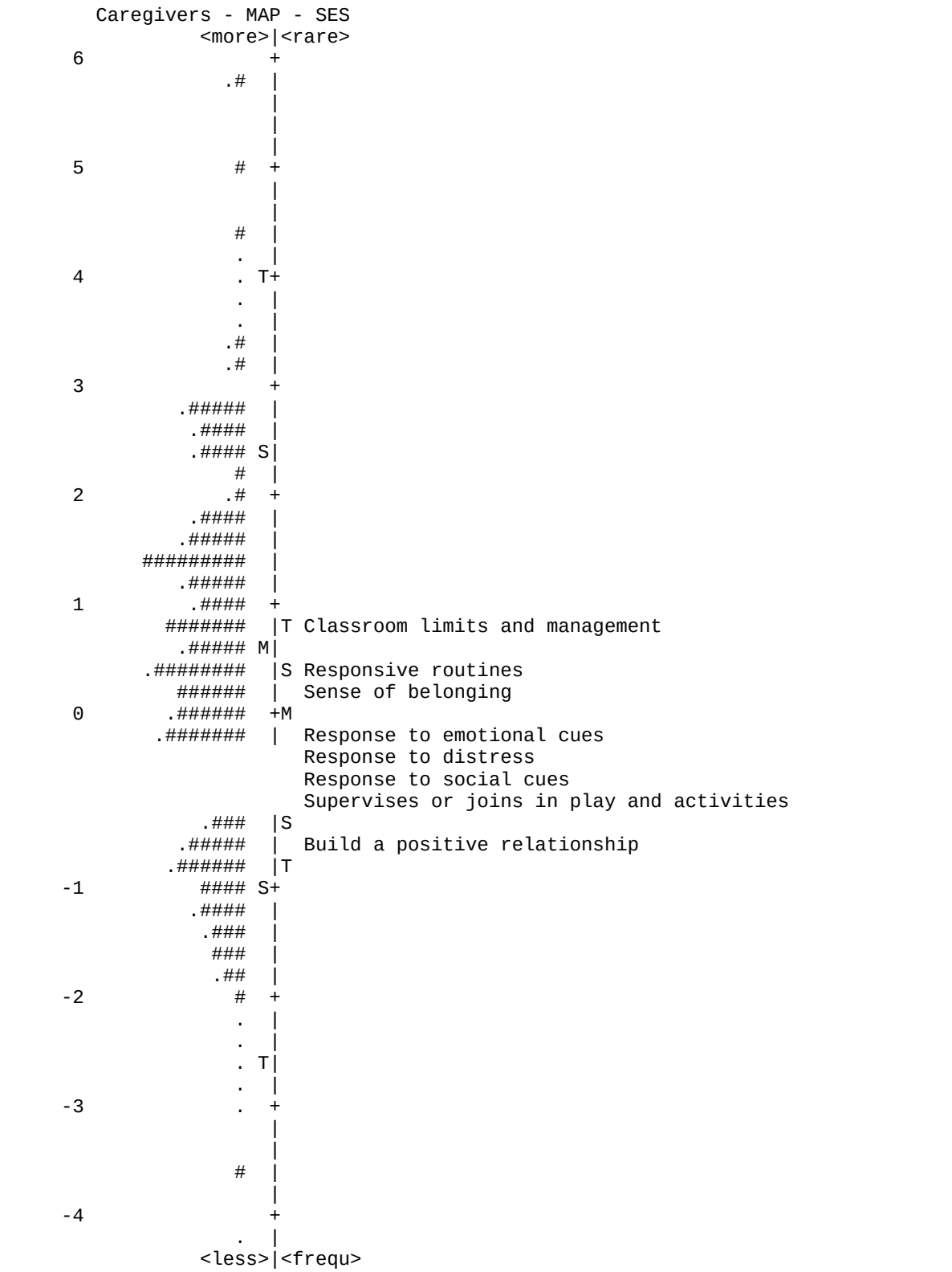


Figure VII.11. Item map for Q-CCIIT support for language and literacy development scale

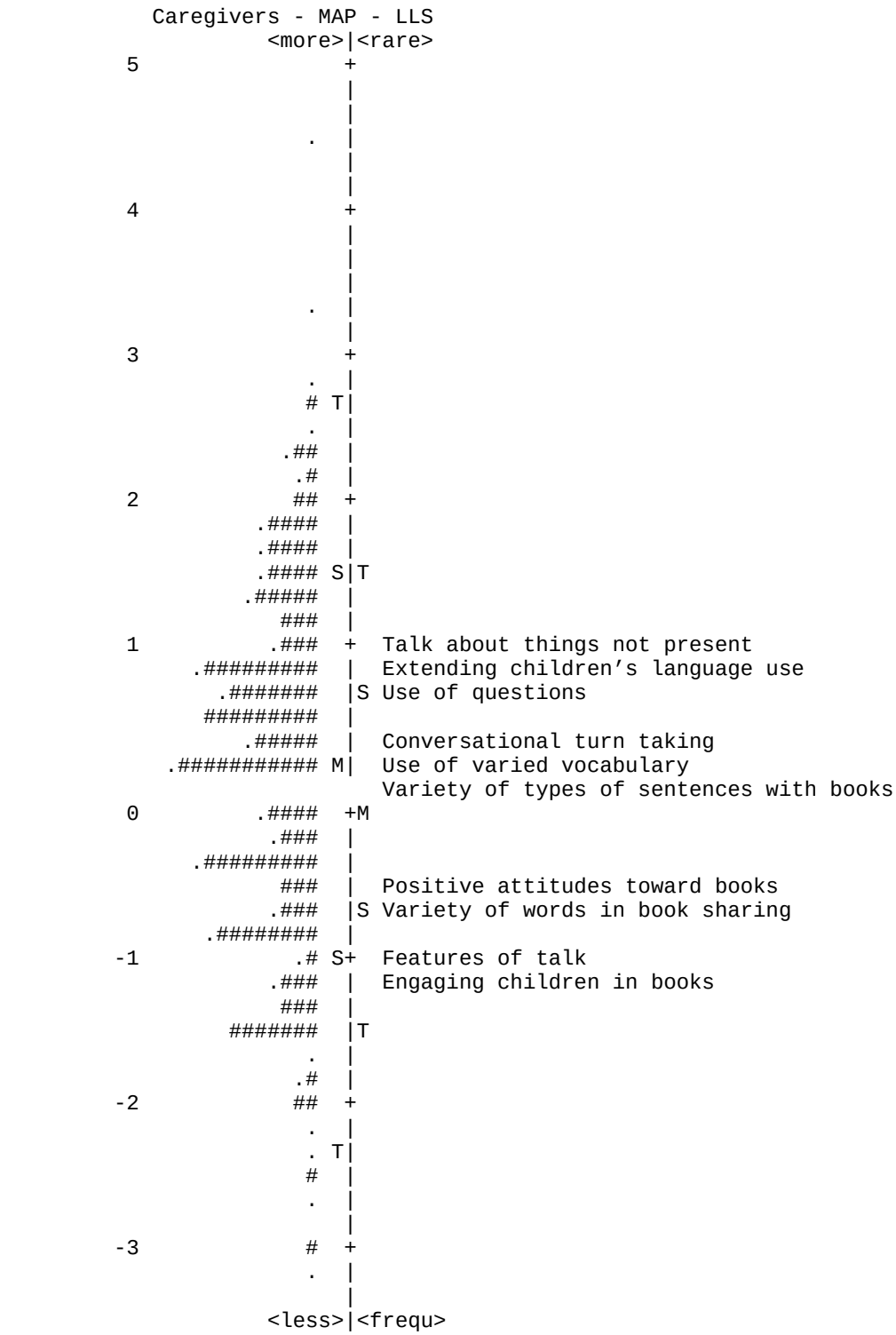
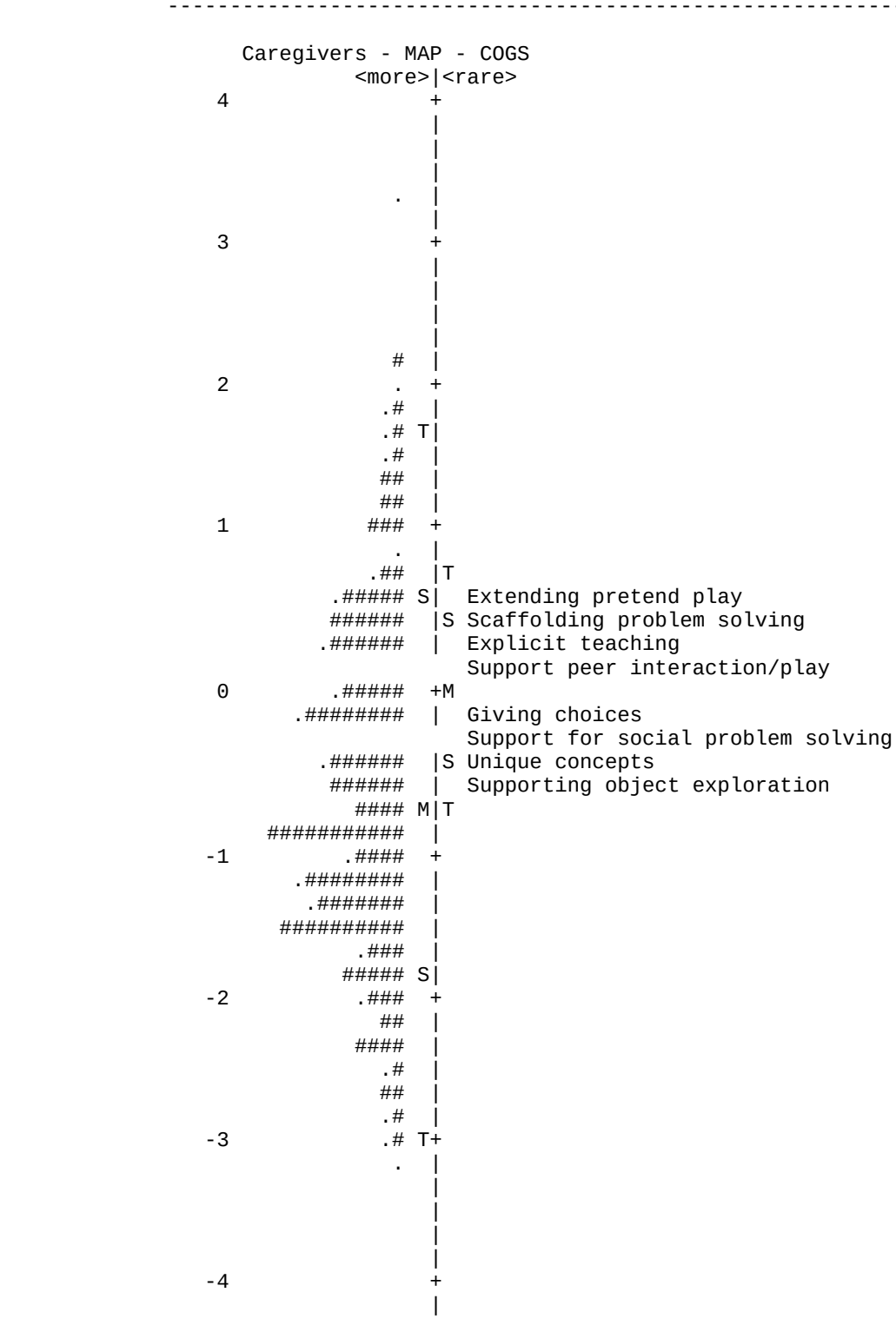


Figure VII.12. Item map for Q-CCIIT support for cognitive development scale



WINSTEPS (software used to estimate Rasch scores) also provides a factor analysis of the residuals to permit an examination of whether any problems with fit to the model are attributable to multidimensionality of the items, thereby providing additional information about how a measure could be strengthened. The results showed that the across-the-visit items and observation cycle items were grouped into separate factors, suggesting method variance in these two types of items. This is consistent with the findings from the CFA and confirms the unidimensionality of the scale content.

We examined classrooms with the most misfit and highest residuals to identify any pattern of misfit for subgroups. We included codes for the classroom setting and family and caregiver characteristics when we created the file for Rasch analysis, appending the codes to the end of the classroom ID so we could identify any patterns when examining the output of the cases with the greatest misfit. For example, if infant classrooms had the most misfit, we would need to look more carefully at the fit of the items for this subgroup to determine whether some should be excluded or redefined for this subgroup.

Summary. Consistent with other reliability estimates, IRT reliability estimates for item difficulty indicated excellent reliability for item response and strong reliability for classroom quality. IRT Rasch analysis supported the construct validity of the Q-CCIIT, with results indicating that the ordering of item difficulties is consistent with theoretical assumptions about whether certain behaviors are more likely to be observed than others for each of the scales. For example, quality support for object exploration was more likely to be observed than quality support for scaffolding problem solving. Overall, the item fit statistics were in the acceptable range, indicating good fit to the model, with the exception of extending pretend play. Extension of pretend play was rarely observed, so estimates are less reliable (that is, some caregivers may score well on representational play although they did not score well on less difficult skills, such as explicit teaching). This difference may reflect caregivers' approach to teaching; for example, teachers using a play based curricular approach may have demonstrated support for pretend play but not explicit teaching. Given its role in early development and our plan to use mean scores rather than Rasch scores, we retained this item. Item maps showed that the models capture different levels of quality, with the three scales showing fairly normal distribution.

K. Differential Item Functioning (DIF) analysis

We conducted DIF analyses in the psychometric field test to evaluate whether the items have the same meaning for different groups, such as program type (center-based classrooms versus FCCs) and child age (infant versus toddler classrooms).[57] We used a separate calibration DIF test with each Q-CCIIT scale and subgroup. Across all subgroups, almost all of the items differed in item difficulty across the subgroups when using t-tests. With large sample sizes and well-estimated parameters, even small differences (differences of 0.04) will be statistically significant; for this reason, we also applied the Mantel method with a logit scale transformation of the Educational Testing Service (ETS) criteria[58] (Zwick, Thayer, and Lewis 1999; Linacre n.d.) to examine DIF parameters, whereby g = $\text{difficulty}_{\text{Group1}}$ - $\text{difficulty}_{\text{Group2}}$, and large

[57] Group sizes were less than 200 for most subgroups, but with Rasch "reasonably robust item difficulties" can be obtained with 30 per group classification (Linacre n.d.)

[58] The ETS criteria use Mantel-Haenszel Delta.

(meaningful) DIF g >= .638, intermediate DIF g >= .426, negligible g < .426. Only large DIF is considered serious enough to remove for a large-scale assessment, and then only after evaluating whether the item has a wording or presentation problem that unfairly limits a particular group. DIF can exist when the behavior assessed by the item is important to the construct, but may differ in frequency across subgroups.

The results generally suggest comparable item difficulties for the Q-CCIIT scales by child age and program type. For Support for Social-Emotional Development (Table VII.15), with regard to child age, none of the items had large DIF when using a separate calibration approach. When using the Mantel method, building a positive relationship had large DIF favoring infant classrooms (less frequently observed in toddler than infant classrooms). In the separate calibration approach, this item (building a positive relationship) had intermediate DIF, as did the item supervises or joins in play and activities (less frequently observed in infant than toddler classrooms). The Mantel test examines the item difficulty, taking into account the classroom quality relative to the other group. Overall, infant classrooms scored lower than toddler classrooms. However, the use of nurturing touch (one of the behaviors needed for a high score on building a positive relationship) is more prevalent in infant than toddler classrooms. Infant classrooms are smaller than toddler classrooms, so it is more common to see evidence of the caregiver's demonstrating relationship-building behaviors with all of the infants in the room. Nurturing touch and working to establish a positive relationship with each child are important across both age groups; therefore, we did not revise the item.

Table VII.15. Comparison of item difficulty: support for social-emotional development

	Child age		Program type		DLL concentration	
	Infant	Toddler	Center	FCC	Low	High
Classroom limits and management[b]	0.86	0.97	0.88	0.84	0.87	0.91
Responsive routines[b]	0.46	0.25	0.36	0.15	0.34	0.24
Sense of belonging[b]	0.22	-0.19	0.04	0.56	0.15	0.27
Responding to emotional cues[a]	-0.19	0.00	-0.10	-0.13	-0.10	-0.14
Supervises or joins in play and activities[b]	0.14	-0.43	-0.11	-0.43	-0.17	-0.27
Responding to social cues[a]	-0.23	-0.23	-0.22	-0.34	-0.22	-0.30
Responding contingently to distress[a]	-0.41	-0.10	-0.29	0.00	-0.23	-0.21
Building a positive relationship[a]	-0.85	-0.27	-0.56	-0.66	-0.63	-0.51

Source: Q-CCIIT Fall 2012 Psychometric Field Test.

Note: Item difficulty based on separate calibrations using a one-parameter Rasch model.

[a] The item rating used in analysis is the mean across valid observation cycles (five minutes or longer), rounded to the nearest integer.

[b] Rated across the visit.

Table VII.15a. Correlations of the difficulty estimates of items between subgroups

	Between infant/toddler classrooms	Between centers/FCCs	Between classrooms with high/low DLLs
Support for Language and Literacy Development	.99	.99	.98
Support for Social-Emotional Development	.71	.84	.98
Support for Cognitive Development	.81	.85	.94

No large DIF was present on the Support for Social-Emotional Development scale by program type, and intermediate DIF was present for only one item (using both methods of examining DIF): sense of belonging, which is observed less frequently in FCCs than in centers. No DIF was present by child age or program type for Support for Language and Literacy Development (Table VII.16) or Support for Cognitive Development (Table VII.17). No DIF was present by DLL concentration for any of the three scales. The absence of large DIF for any subgroup (other than the infant-toddler difference in building a positive relationship) indicates that the Q-CCIIT can be used with diverse classrooms. Due to the environment differences, the environment aspects of the sense of belonging item (for example, providing a space for each child's personal belongings) may be unfair to FCCs.

Table VII.16. Comparison of item difficulty: support for language and literacy development

	Child age		Program type		DLL concentration	
	Infant	Toddler	Center	FCC	Low	High
Talk about things not present[b]	1.09	1.08	1.07	1.05	1.08	1.07
Extending children's language use[a]	0.86	1.00	0.93	0.79	0.90	0.90
Use of questions[a]	0.69	0.76	0.73	0.63	0.75	0.57
Conversational turn taking[a]	0.42	0.32	0.37	0.32	0.44	0.21
Caregiver use of varied vocabulary[a]	-0.01	0.29	0.15	0.36	0.20	0.18
Variety of types of sentences (book sharing)[a]	0.08	0.13	0.11	0.05	0.10	0.12
Positive attitude toward books[b]	-0.40	-0.61	-0.51	-0.52	-0.59	-0.36
Variety of words (book sharing)[a]	-0.59	-0.63	-0.61	-0.67	-0.65	-0.60
Features of talk[b]	-1.01	-1.17	-1.08	-0.86	-1.15	-0.85
Engaging children in books[a]	-1.11	-1.18	-1.15	-1.14	-1.09	-1.25

Source: Q-CCIIT Fall 2012 Psychometric Field Test.

Note: Item difficulty based on separate calibrations using a one-parameter Rasch model.

[a]The item rating used in analysis is the mean across valid observation cycles (five minutes or longer), rounded to the nearest integer.

[b] Rated across the visit.

Table VII.17. Comparison of item difficulty: support for cognitive development

	Child Age		Program Type		DLL Concentration	
	Infant	Toddler	Center	FCC	Low	High
Extending pretend play[b]	0.74	0.39	0.51	0.24	0.51	0.27
Scaffolding problem solving[a]	0.38	0.34	0.36	0.46	0.39	0.38
Explicit teaching[b]	0.28	0.21	0.24	0.06	0.17	0.30
Supporting peer interaction/play[a]	-0.16	0.23	0.08	0.12	0.07	0.12
Support for social problem solving[b]	0.06	-0.18	-0.11	-0.11	-0.08	-0.17
Giving choices[b]	0.06	-0.24	-0.12	-0.01	-0.10	-0.09
Unique concepts	-0.42	-0.32	-0.35	-0.51	-0.44	-0.28
Supporting object exploration[a]	-0.93	-0.43	-0.62	-0.26	-0.51	-0.53

Source: Q-CCIIT Fall 2012 Psychometric Field Test.

Note: Item difficulty based on separate calibrations using a one-parameter Rasch model.

[a] The item rating used in analysis is the mean across valid observation cycles (five minutes or longer), rounded to the nearest integer.

[b] Rated across the visit.

We examined correlations of the estimates of the difficulty of the items based on separate calibrations by subgroups (Table VII.15a). The correlations of the item difficulty estimates were generally high. With two items showing strong to moderate DIF between infants and toddlers, the correlation of the social-emotional scale item difficulties (N=8) for infants versus toddlers was the weakest (r = 0.71).

We also examined the correlations between the classroom scores based on separate calibration IRT estimates and the mean raw scores (mean ratings). The correlations for all subgroups and scales were greater than 0.97 and most were greater than 0.99.

Summary. DIF analyses generally suggested comparable item difficulties for the Q-CCIIT scales by child age, program type, and concentration of DLLs. None of the positive scales had large DIF by child age or program type, and two had no DIF. There was no DIF by DLL concentration for any of the three scales. The absence of any large DIF for any subgroup suggests that the Q-CCIIT can be used with diverse classrooms.

L. Assessing convergent and discriminant validity

The psychometric field test included analysis of the QCCIIT's validity using a range of methods. First, the Technical Working Group (TWG) members reviewed the items and provided the initial feedback on the content validity of the items and proposed scales. Second, as noted above, the examination of the item difficulty hierarchy in the IRT analysis and the estimates of factor loadings in the CFAs provided evidence of construct validity. Finally, we examined associations with the concurrent observations using the Observational Record of the Caregiving Environment (ORCE; in all settings and the Infant/Toddler Environment Rating Scale-Revised (ITERS-R) or Family Child Care Environment Rating Scale-Revised (FCCERS-R) depending on

setting type, as well as with classroom and caregiver characteristics (caregiver-child ratio, class size, and caregiver education and experience). We examined concurrent validity by program type (centers and FCCs) and child age (infant and toddler classrooms). Since most of the FCCs are mixed-age groups, the analyses by child age are limited to infant and toddler classrooms in centers. This section presents information on convergent and discriminant validity for these measures.

Convergent Validity with the ORCE. The ORCE had higher scores for infant classrooms than for toddler classrooms and for FCCs than toddler classrooms, both for the overall ORCE Overall Qualitative Ratings and most of the other ORCE scores (Table VI.3). This finding is inconsistent with the findings on both the environmental rating scales (ITERS-R and FCCERS-R) (Tables VI.4 and VI.5) and the Q-CCIIT (Tables VII.1). The difference in observation method—the ORCE focuses on individual children rather than on the average experience of the children in the room—is likely at least one of the reasons for these differences. The difference in observation method probably also affected the magnitude of the correlations of the ORCE with the Q-CCIIT. What is not clear is whether the observation method focusing on the individual child provides a more obvious cue about the focus of the observer attention influencing caregiver attention in infant classrooms and FCCs than the busy toddler rooms, or if the items are more sensitive to infant classrooms. Greater variance was evident in the ORCE Overall Qualitative Rating in infant classrooms and FCCs than in toddler classrooms. Many of the other scores on the ORCE with the full sample and the subgroups had limited variance, potentially making it more difficult to detect correlations (Tables VI.2 and VI.3).

Despite these challenges, the ORCE Overall Qualitative Rating and almost all of the scores were significantly correlated with the Q-CCIIT scales (Table VII.18.) The strongest relationships were found with Support for Social-Emotional Development scale. This finding is consistent with the theoretical/conceptual framework for these two measures and provides evidence of convergent validity. The Q-CCIIT Support for Cognitive Development scale had the fewest relationships with the ORCE, providing some evidence for the discriminant validity of the scales (Table VII.18). In addition, the scales in the ORCE that focused on negative behaviors (later reverse coded) were more strongly related to the Q-CCIIT Areas of Concern and Extreme Concern than to the positive scales. Similarly, the positive scales on the ORCE were usually more strongly related to the Q-CCIIT positive scales than to the Areas of Concern and Extreme Concern.

Despite the greater variance on the ORCE infant classroom scores, the correlations of the ORCE Overall Qualitative Rating with the Q-CCIIT Support for Language and Literacy Development and Support for Cognitive Development scales were stronger for toddler than infant classrooms and FCCs (Appendix G, Tables G.3, G.6, and G.8).

Convergent and Discriminant Validity with Environmental Rating Scales: ITERS-R and FCCERS-R. Across both center-based classrooms (ITERS-R) and FCCs (FCCERS-R), we found convergent and discriminant validity evidence for the Q-CCIIT. Similar to the Q-CCIIT, toddler classrooms had higher scores on the ITERS-R than infant classrooms (Table VI.5) or than the FCC scores on the FCCERS-R, with the exception of Personal Care and Social Interaction (Table VI.4). On the latter two scales, infant classrooms scored higher than toddler classrooms or FCCs.

The ITERS-R Total score was correlated with each of the Q-CCIIT scales, with the strongest correlation with Support for Social-Emotional Development ($r = 0.54$), followed by Support for Language and Literacy Development ($r = 0.36$), and Support for Cognitive Development ($r = 0.31$). the strongest correlations between the ITERS-R and Q-CCIIT scales were between the Q-CCIIT positive scales and the ITERS-R Listening and Talking scale (Table VII.19). The FCCERS-R Total score was also correlated with each of the Q-CCIIT scales (Table VII.20), although the strongest correlation was with the Support for Cognitive Development ($r = 0.50$), rather than with Support for Social-Emotional Development ($r = 0.39$).

Some evidence for discriminant validity was found on the ITERS-R and FCCERS-R. In the full sample, no relationship was found between the FCCERS-R Personal Care or Space and Furnishings and any Q-CCIIT scale. No relationship was found for the ITERS-R Personal Care scale with Support for Language and Literacy Development or Support for Cognitive Development. The ITERS-R Space and Furnishings scale did have a relationship with Support for Language and Literacy Development in the full sample, but no relationship with the other Q-CCIIT scales (Table VII.19). Next, we discuss separately subgroup differences in relationships of the ITERS-R and FCCERS-R for each of the Q-CCIIT scales.

Support for Social-Emotional Development. Support for Social-Emotional Development was positively correlated with all ITERS-R subscales in center-based classrooms. For FCC settings, FCCERS-R Listening and Talking, Social Interaction, Activities, and Program Structure subscales all showed positive correlations with the Support for Social-Emotional Development scale. The Program Structure items reflect factors that may be needed to be most supportive of social-emotional development (for example, flexibility and responsiveness of schedules and activities) and that address areas similar to the across-the-visit items on the Q-CCIIT Support for Social-Emotional Development scale (for example, responsive routines).

Table VII.18. Correlations between Q-CCIIT scales and the ORCE scales

	Q-CCIIT Scales				
ORCE Scales	Support for social-emotional development	Support for language and literacy development	Support for Cognitive development	Areas of concern	Extreme concern
Overall Qualitative Rating	0.50***	0.35***	0.39***	-0.22*	-0.46***
Sensitivity/Responsiveness to Distress [a]	0.08	-0.10	0.08	-0.06	-0.24*
Sensitivity/Responsiveness to Non-Distress	0.53***	0.40***	0.42***	-0.37***	-0.47***
Lack of Intrusiveness	0.39***	0.26**	0.28**	-0.42***	-0.51***
Lack of Detachment/Disengagement	0.36***	0.27**	0.26**	-0.08	-0.37***
Stimulation of Cognitive Development	0.43***	0.36***	0.38***	-0.18	-0.28**
Positive Regard for the Child	0.39***	0.29**	0.33***	-0.25**	-0.46***
Lack of Negative Regard for the Child	0.20*	0.11	0.13	-0.30***	-0.32***
Lack of Flatness of Affect	0.27**	0.20*	0.18	-0.00	-0.26**
Fostering Exploration[b]	0.44**	0.33*	0.34*	-0.22	-0.38*
Positive Rating	0.50***	0.36***	0.41***	-0.29*	-0.44***
Lack of Negative Rating	0.40***	0.29**	0.28**	-0.18**	-0.44***
Language Stimulation	0.30***	0.28**	0.25**	-0.04*	-0.22*
Positive Behavior Toward Child	0.14	0.01	0.01	-0.09	-0.19*
Negative Behavior Toward Child	-0.36***	-0.40***	-0.44***	0.26**	0.19*
Sample Size	**40–119**	**40–119**	**40–119**	**40–119**	**40–119**

Source: Q-CCIIT Fall 2012 Psychometric Field Test Data.

*p<.05; **p<.01; ***p<.001.

ORCE = Observational Ratings of the Caregiving Environment.

Table VII.19. Correlations between Q-CCIIT scales and ITERS-R subscales and child-adult ratio

	Q-CCIIT scales				
ITERS-R subscales	Support for social-emotional development	Support for language and literacy development	Support for cognitive development	Areas of concern	Extreme concern
ITERS-R Total	**0.54*** **	**0.36** **	**0.31***	**-0.10**	**-0.33** **
Listening and Talking	0.57***	0.40**	0.41***	0.06	-0.43***
Social Interaction	0.47***	0.22	0.25	-0.08	-0.46***
Activities	0.41***	0.29*	0.30*	-0.13	-0.20
Program Structure	0.54***	0.35**	0.24	-0.23	-0.34**
Space and Furnishings	0.27*	0.32*	0.20	-0.08	0.01
Personal Care	0.26*	0.13	0.01	-0.01	-0.07
Child/Adult Ratio	-0.23	-0.05	-0.14	-0.12	0.31*
Sample Size	**63–64**	**63–64**	**63–64**	**62–63**	**62–63**

Source: Q-CCIIT Fall 2012 Psychometric Field Test Data.

*p<.05; **p<.01; ***p<.001.

ITERS-R = Infant-Toddler Environment Rating Scale-Revised.

Support for Language and Literacy Development. The Q-CCIIT Support for Language and Literacy Development scale is positively correlated with total ITERS-R and FCCERS-R quality scores for both center-based and FCC as well as toddler (but not infant) classrooms. Support for Language and Literacy Development was positively correlated with the ERS Listening and Talking subscale, though this correlation was not significant when looking only at infant classrooms. This could be due to more limited variation on the Listening and Talking scales in classrooms with preverbal infants. In infant classrooms, Social Interaction was the only ITERS-R subscale that was positively correlated with Language and Literacy Development. The Social Interaction subscale might capture those interactions most supportive of preverbal infants' language development. Program Structure is correlated with Support for Language and Literacy Development in centers and FCC settings as well as toddler (but not infant) classrooms. It could be that the Program Structure subscale captures the stability and responsiveness that undergird more positive caregiving.

Table VII.20. Correlations between Q-CCIIT scales and FCCERS-R subscales and child-adult ratio

	Q-CCIIT scales				
FCCERS-R subscales	Support for social-emotional development	Support for language and literacy development	Support for cognitive development	Areas of concern	Extreme concern
FCCERS-R Total	**0.39****	**0.29***	**0.50*****	**-0.36***	**-0.45****
Listening and Talking	0.42**	0.32*	0.43**	-0.35*	-0.35*
Social Interaction	0.39**	0.21	0.43**	-0.41**	-0.49***
Activities	0.33*	0.21	0.46**	-0.18	-0.23
Program Structure	0.33*	0.32*	0.41**	-0.23	-0.40**
Space and Furnishings	0.11	0.15	0.28	-0.28*	-0.31*
Personal Care	0.19	0.21	0.22	-0.20	-0.32*
Child/Adult Ratio	-0.03	0.14	-0.01	0.05	0.31*
Sample Size	**48–49**	**48–49**	**48–49**	**48–49**	**48–49**

Source: Q-CCIIT Fall 2012 Psychometric Field Test Data.

*p<.05; **p<.01; ***p<.001.

FCCERS-R = Family-Child Care Environment Rating Scale-Revised.

Support for Cognitive Development. The ITERS-R and FCCERS-R Total scores are correlated with Support for Cognitive Development in centers and FCC settings. Among the subscales, the Activities scale, Social Interaction, and Program Structure had correlations in one or more subgroups. The ITERS-R and FCCERS-R Activities subscales were positively correlated with Support for Cognitive Development in centers and FCCs, as well as toddler classrooms. The variety and quality of activities allows more opportunities to explore and scaffold learning, and this relationship was evident in the toddler but not the infant classrooms, where lower average Activities scores might have lessened our ability to observe any meaningful correlation. In infant classrooms, the ITERS-R Social Interaction scale was moderately related to Support for Cognitive Development; both scales include items focused on peer interaction. In FCCs, in addition to the Activities subscale, Program Structure was correlated with Support for Cognitive Development.

Areas of Concern and Extreme Concern. Overall, infant classrooms score lower than toddler classrooms on the Q-CCIIT. The Areas of Concern scale in infant and toddler classrooms is so low that any correlations found may be spurious. In FCCs, the FCCERS-R Listening and Talking and Social Interaction scores were significantly and negatively correlated with Areas of Concern. Although Areas of Concern and Listening and Talking do not align conceptually, it is likely that settings with high Listening and Talking scores are less likely to exhibit Areas of Concern because they are of higher overall quality. Program Structure was significantly and negatively correlated with Areas of Concern for FCCs. Areas of Concern was significantly and negatively correlated with FCCERS-R Total scores.

One of the scoring methods on the ITERS-R and the FCCERS-R requires that classrooms receive a score of one if any of the specified negative practices was observed. This may have contributed to stronger negative relationships between scores on the ITERS-R/FCCERS-R and the Q-CCIIT Areas of Concern. In some cases, the correlation with Areas of Concern was stronger than the correlation with the positive Q-CCIIT scales. It suggests that the ERS scoring method weakened the estimated relationship of the ITERS-R and FCCERS-R with the positive Q-CCIIT scales. Similarly, after examining the anchors on some of the items, we hypothesize that the correlation between scales measuring similar constructs might have been stronger (for example, Listening and Talking with Support for Language and Literacy Development) if scoring on the ITERS-R and FCCERS-R did not require meeting all the criteria of a previous rating level before moving to the next level.

Correlations between Extreme Concern and ERS measures could not be estimated reliably in infant and toddler classrooms because the incidence was extremely low. In the FCC sample, the FCCERS-R Total score as well as the Listening and Talking, Social Interaction, Space and Furnishings, Personal Care, and Program Structure subscales were negatively correlated with Extreme Concern.

Q-CCIIT Associations with Validation Measures. With high inter-factor correlations on the Q-CCIIT, we also used ordinary least squares regression to examine the contribution of each of the scales in explaining shared variance with the ORCE, ITERS-R and FCCERS-R (Table VII.21). The Q-CCIIT explained about one-third of the variance in each of the validation measures (R^2 = 0.29, 0.33, 0.32, respectively). For both the ORCE and the ITERS-R, the Q-CCIIT Support for Social-Emotional Development explained the majority of the shared variance. The ORCE and the FCCERS-R samples both included FCCs and for those samples the Areas of Concern added at least marginally to the explanation of the variance after controlling for the other Q-CCIIT scales. The Support for Cognitive Development rather than Support for Social-Emotional Development explained the majority of the variance on the FCCERS-R.

Although the ITERS-R and FCCERS-R address similar constructs, the difference noted in the associations with the Q-CCIIT may be related to differences in how these constructs are measured. The cognitive demand is stronger on the FCCERS-R items than on the ITERS-R items, likely due to the broad age range addressed on the FCCERS-R. For example, for ratings of 7 on the item about helping children use language, the FCCERS-R requires providers to ask questions that "encourage more complex answers" while the ITERS-R only requires that staff "ask children simple questions." At the low end of the scale, a rating of 3 on the use of books item requires daily use of books with children on the FCCERS-R while the ITERS-R only requires the use of books 3 times a week. The FCCERS-R has an item about math while the ITERS-R does not.

Table VII.21. Associations between Q-CCIIT and validation scales: OLS results

	ORCE qualitative item mean	ITERS-R total score	FCCERS-R total score
Support for Social Emotional Development	.372*	.547 **	-.133
Support for Language and Literacy	-.127	.059	-.074
Support for Cognitive Development	.111	-.114	.515 *
Areas of Concern	-.233*	-.099	-.309 †
R^2	.292	.328	.323

Source: Q-CCIIT Fall 2012 Psychometric Field Test
Note: Model results reported as Beta coefficients.
†p<.10; *p<.05. **p<.01; ***p≤.001.

Summary. Overall, convergent validity with the Q-CCIIT was evident, with expected moderate to high moderate relationships found with the ORCE, FCCERS-R, and ITERS-R; evidence of discriminant validity was also found, with weaker or no relationships between the Q-CCIIT and subscales such as the ITERS-R Personal Care subscale. First, we found some evidence of convergent and discriminant validity evidence for the Q-CCIIT using the ORCE. With regard to convergent validity, the ORCE Overall Qualitative Rating and almost all of the scores showed a relationship with the Q-CCIIT scales. However, some challenges to convergent validity arose. For example, the descriptive statistics by subgroup showed that the ORCE scores were generally higher for infant classrooms and FCCs than toddler classrooms, perhaps due to a difference in observation method between the Q-CCIIT and the ORCE. In addition, there was greater variance in the ORCE Overall Qualitative Rating in infant classrooms and FCCs than in toddler classrooms; however, the correlations of the ORCE Overall Qualitative Rating with the Q-CCIIT Support for Language and Literacy Development and Support for Cognitive Development scales were stronger for toddler classrooms than for infant classrooms and FCCs. Many of the other scores on the ORCE with the full sample and the subgroups had limited variance, potentially making it more difficult to detect correlations. We also found some evidence for discriminant validity of the scales, as the Q-CCIIT Support for Cognitive Development had the fewest relationships with the ORCE. The scales in the ORCE that focused on negative behaviors (later reverse coded) were more strongly related to the Q-CCIIT Areas of Concern and Extreme Concern than to the positive scales. Similarly, the positive scales on the ORCE were usually related more strongly to the Q-CCIIT positive scales than to the Areas of Concern and Extreme Concern.

Second, we found convergent and discriminant validity evidence for the Q-CCIIT across both center-based classrooms (using ITERS-R) and FCCs (using FCCERS-R). With regard to convergent validity, similar to the Q-CCIIT, toddler classrooms had higher scores on the ITERS-R than infant classrooms or than the FCCERS-R scores, with the exception of Personal Care and Social Interaction. Both the ITERS-R Total score and the FCCERS-R Total score were correlated with each of the Q-CCIIT scales. We also found some evidence for discriminant validity on the ITERS-R and FCCERS-R. Specifically, in the full sample, no relationship was found between the FCCERS-R Personal Care or Space and Furnishings subscales and any Q-

CCIIT scale. No relationship was found for the ITERS-R Personal Care subscale with Support for Language and Literacy Development or Support for Cognitive Development. The ITERS-R Space and Furnishings did have a relationship with Support for Language and Literacy Development in the full sample, but no relationship with the other Q-CCIIT scales.

M. Validity evidence: caregiver characteristics and ratio

We examined relationships between the Q-CCIIT scales and caregiver characteristics, including education, experience, and caregiver reports of depressive symptoms. For education, we constructed variables that captured the mean number of caregivers with an associates' degree or higher and one that captured whether any of the caregivers in the classroom had at least an associate's degree. We looked at correlations with each of the Q-CCIIT mean scale scores and Areas of Concern and the number of Areas of Extreme Concern. With so many comparisons and few significant findings, we report only significant results. The significance level has not been adjusted for the number of correlations tested, and some findings may be spurious.

Support for Social-Emotional Development was associated with caregiver education. Classrooms in which caregivers had on average an associate's degree or higher provided stronger Support for Social-Emotional Development in the overall sample ($r = 0.14$, $p<0.01$), in center-based care ($r = 0.17$, $p<0.01$), and in toddler classrooms ($r = 0.20$, $p<0.05$).

In the overall sample, Support for Language and Literacy Development was associated with having a caregiver who plans to work in a child care setting the following year ($r = 0.10$, $p<0.05$). No other caregiver characteristics correlated significantly with the language scale.

With Support for Cognitive Development, we found a small negative correlation between mean years spent working with infants and toddlers for both the full sample and FCC settings ($r = -0.11$, $p<0.05$ and $r = -0.20$, $p<0.05$, respectively).

In infant classrooms, the highest number of infant development courses taken by any one of the classroom caregivers had a positive correlation with Support for Cognitive Development ($r = 0.20$, $p<0.05$), as did a center caregiver's having an associate's degree ($r = 0.13$, $p<0.05$). Having a moderately to highly depressed caregiver in a toddler classroom was negatively related with Support for Cognitive Development ($r = -0.20$, $p <0.05$).

Child/adult ratios were not related to any of the Q-CCIIT scales. With licensing rules in place, the child/adult ratio was positive in most classrooms and thus did not relate to the variation in classroom quality. The limited findings in relation to caregiver education and experience associations with positive caregiving are consistent with other research that found that observed classroom quality—both interaction and environment quality—predicted child outcomes, while other structural quality indicators used in quality rating and improvement systems (QRIS) were weak or not significant predictors (Sabol et al. 2013).

Even with the limited variance in the Areas of Concern, we found relationships with some caregiver characteristics. Mentoring and education were negatively correlated with Areas of Concern, but these correlations were not consistent across variables and subgroups. For example, mentoring ($r = -0.14$, $p<0.01$) and the number of early childhood education courses ($r = -0.13$, $p<0.05$) were negatively correlated with Areas of Concern in the overall sample. Correlations for

these variables were in the same direction in each of the subsamples and of similar magnitude for infant and toddler classrooms, but did not reach significance, likely due to lack of power.

The mean number of caregivers with an associate's or higher degree was negatively related to Extreme Concern for the full sample (r = -0.13, p<0.05) and to Areas of Concern, for centers overall (r = -0.20, p<0.001), infants (r = -0.24, p<0.01), and toddlers (r = -0.18, p<0.05), but not for FCCs.

We were surprised to find some positive associations between experience and Areas of Concern. In the full sample (r = 0.19, p<0.001), toddler classrooms (r = 0.19, p<0.001), and FCCs (r = 0.24, p<0.05), the classroom average of caregivers' experience in working with infants and toddlers was associated with Areas of Concern. Similarly, the greatest number of years of experience with infants and toddlers of any teacher in the classroom was positively associated with Areas of Concern (r = 0.12, p<0.05 and r = 0.21, p<0.05 for full and FCC samples, respectively). It is likely that the overall sample estimates were driven mainly by the FCCs, given the combination of both more limited variance in centers on Areas of Concern and a greater number of caregivers present in infant and toddler classrooms.

Areas of Concern was not significantly correlated with child/adult ratio, although higher child/adult ratios in FCCs were significantly associated with greater Extreme Concern scores (r = 0.31, p<0.05).

Mentoring (r = -0.14, p<0.01), education (caregiver having an associate's degree, r = -0.13, p<0.05), ECE coursework (r = -0.15, p<0.01), and child (r = -0.13, p<0.05) and infant (r = -0.12, p<0.05) development courses all were negatively related to Extreme Concern in the full sample.

In FCCs, caregiver reports of moderate to high levels of depressive symptoms were positively correlated with Extreme Concern (r = 0.24, p<0.05), though not in any other settings. Given the greater number of caregivers in center-based classrooms and the fact that the Q-CCIIT score is based on the average experiences of children across caregivers, it is not surprising that a correlation was not found in center-based settings.

Summary. Caregiver characteristics had a weak relationship with the three positive Q-CCIIT scales, and child/adult ratios were not related to any of the Q-CCIIT scales. Even with the limited variance in the Areas of Concern, we found relationships with some caregiver characteristics; correlations for these variables were in the same direction in each of the subsamples, and of similar magnitude for infant and toddler classrooms, but did not reach significance. Somewhat surprisingly, we found some positive associations between experience and Areas of Concern; it is likely that the overall sample estimates are driven mainly by the FCCs, given the combination of both more limited variance in centers on Areas of Concern and a greater number of caregivers present in infant and toddler classrooms. Areas of Concern scores were not significantly correlated with child/adult ratio, although higher child/adult ratios in FCCs were significantly associated with greater Extreme Concern scores (r = 0.31). In FCCs, caregiver reports of moderate to high levels of depressive symptoms were positively correlated with Extreme Concern (r = 0.24), although not in any other settings. These correlations are the highest among the correlations we examined.

N. Overall summary

These analyses provide psychometric evidence supporting the reliability and validity of the Q-CCIIT as a measure of caregiving quality. Adequate to strong reliability was found across multiple analytic methods. Specifically, the Q-CCIIT measure demonstrated the following:

- sufficient internal consistency reliability, with strong reliability for the positive scales and weaker, though acceptable, reliability for the Areas of Concern;
- adequate temporal stability for the overall sample as well as for infant classrooms and FCCs;
- positive inter-rater reliability, with slightly lower estimates for FCCs;
- and excellent IRT reliability estimates for item response and strong reliability for classroom quality.

The G-study indicated that most of the variance may be attributed to differences in classrooms and the interaction of classrooms with items and cycles. D-study results indicated that the G-coefficient and dependability index (phi) for each of the three positive scales showed a good level of reliability.

Both CFA and IRT Rasch analyses supported the construct validity of the Q-CCIIT. DIF analyses generally suggested comparable item difficulties for the Q-CCIIT scales by child age, program type, and concentration of DLLs.

Convergent validity with related measures was evident, with expected moderate to high-moderate relationships found with the ORCE, FCCERS-R, and ITERS-R. We also found some evidence for discriminant validity of the scales, as the Q-CCIIT Support for Cognitive Development had the fewest relationships with the ORCE, suggesting that the Q-CCIIT is measuring those aspects of interaction not captured in the ORCE. Evidence of discriminant validity was also found, with weaker or no relationships between the Q-CCIIT and scales such as the ITERS-R Personal Care subscale. Caregiver characteristics had a weak relationship with Q-CCIIT scales, and child/adult ratios were not related to any of the Q-CCIIT scales.

This page has been left blank for double-sided copying.

Allen, M. J., & Yen, W. M. (2002*). Introduction to Measurement Theory. Long Grove, IL: Waveland Press.*

Als, H., F. H. Duffy, G. B. McAnulty, M. J. Rivkin, S. Vajapeyam, R. V. Mulkern, S. K. Warfield, P. Huppi, S. C. Butler, N. Conneman, C. Fischer, and E. C. Eichenwald. "Early Experience Alters Brain Function and Structure." *Pediatrics,* vol., 113 no. 4, 2004, pp. 846–857.

American Academy of Pediatrics, American Public Health Association, and National Resource Center for Health and Safety in Child Care and Early Education. "Stepping Stones to Caring for Our Children: National Health and Safety Performance Standards; Guidelines for Early Care and Education Programs, Third Edition." Elk Grove Village, IL: American Academy of Pediatrics, American Public Health Association, and National Resource Center for Health and Safety in Child Care and Early Education, 2013. Available at [http://nrckids.org/default/assets/File/SteppingStones3v4.pdf]. Accessed October 28, 2013.

Atkins-Burnett, S., Monahan, S., Tarullo, L., Barrios, V., E. Cavadel, F. Hurwitz, A. K. Klein, L. Malone and M. Putnam. "Measuring the Quality of Caregiver-Child Interactions with Infants and Toddlers: The Q-CCIIT User's Guide." Princeton, NJ: Mathematica Policy Research, 2014.

Ayoub, C., C.D. Vallotton, and A.M. Mastergeorge. "Developmental Pathways to Integrated Social Skills: The Roles of Parenting and Early Intervention." *Child Development,* vol. 82, 2011, pp. 583–600.

Barrios, V., E. Cavadel, F. Hurwitz, A. K. Klein, L. Malone, S. Monahan, M. Putnam, and R. Weiner. "Q-CCIIT Manual." Princeton, NJ: Mathematica Policy Research, August 2012.

Bernstein, V. J., E. J. Harris, C. W. Long, E. Iida, and S. L. Hans. "Issues in the Multi-Cultural Assessment of Parent–Child Interaction: An Exploratory Study From the Starting Early Starting Smart Collaboration." *Journal of Applied Developmental Psychology,* vol. 26, no. 3, 2005, pp. 241–275.

Booth, A. "Causal Supports for Early Word Learning." *Child Development,* vol. 80, no. 4, 2009, pp. 1243–1250.

Bornstein, M. H., and C. S. Tamis-LeMonda. "Maternal Responsiveness and Cognitive Development in Children.**"** New Directions for Child Development, vol. 43, 1989, pp. 49**–**61**.**

Bovey, T., and P. Strain. "Using Classroom Activities and Routines as Opportunities to Support Peer Interaction." CSEFEL What Works Brief **#5**. Available online at http://csefel.vanderbilt.edu/briefs/wwb5.pdf, Accessed September 23, 2011.

Bradley, R. H., B. M. Caldwell, and R. F. Corwyn. "The Child Care HOME Inventories: Assessing the Quality of Family Child Care Homes." *Early Childhood Research Quarterly*, vol. 18, 2003, pp. 294–309.

Brennan, R. L. (2000). "Performance assessments from the perspective of generalizability theory." *Applied Psychological Measurement*, 24, 339–353.

Brennan, R. L. (2001a). *Generalizability Theory*. Springer-Verlag.

Brennan, R. L. (2001b). Manual for urGENOVA. Iowa City, IA: Iowa Testing Programs, University of Iowa. Version 2.1

Brown, Ari. "Media Use by Children Younger Than 2 Years." *Pediatrics*, vol. 128, no. 5, 2011, pp. 1040–1045.

Brown, T.A. *Confirmatory Factor Analysis for Applied Research*. New York: Guilford Press, 2006.

Caldwell, B., and R. Bradley. *Home Observation for Measurement of the Environment (HOME) Revised Edition.* Little Rock, AR: University of Arkansas, 1984.

Carlson, F. M., B. G. Nelson. "Reducing Aggression with Touch." *Dimensions of Early Childhood*, vol. 34, no. 3, 2006, p. 9-15.

Chan, S., and E. Lee. "Families with Asian Roots." In *Developing Cross Cultural Competence*, edited by E. W. Lynch and M. J. Hanson. Baltimore, MD: Paul H. Brooks Publishing, 2004.

Cook, R.E., M. D. Klein, and A. Tessier. *Adapting Early Childhood Curricula for Children in Inclusive Settings, Sixth edition*. Upper Saddle River, NJ: Pearson, 2004.

Center on the Developing Child at Harvard University. "A Science-Based Framework for Early Childhood Policy: Using Evidence to Improve Outcomes in Learning, Behavior, and Health for Vulnerable Children." Cambridge, MA: Center on the Developing Child at Harvard University, 2007. Available at http://developingchild.harvard.edu/index.php/resources/reports_and_working_papers/policy_framework/.

Certain, Laura K., and Robert S. Kahn. "Prevalence, Correlates, and Trajectory of Television Viewing Among Infants and Toddlers." *Pediatrics*, vol. 109, no. 4, 2002, pp. 634–642.

Duch, Helena, Elisa M. Fisher, Ipek Ensari, Marta Font, Alison Harrington, Caroline Taromino, Jonathan Yip, and Carmen Rodriguez. "Association of Screen Time Use and Language Development in Hispanic Toddlers: A Cross-Sectional and Longitudinal Study." *Clinical Pediatrics*, vol. 52, no. 9, 2013, pp. 857–865.

De Wolff, M. S., and M. H. van Ijzendoorn. "Sensitivity and Attachment: A Meta-Analysis on Parental Antecedents of Infant Attachment." *Child Development*, vol. 68, 1997, pp. 571–591.

Dodici, B. J., D. C. Draper, and C. A. Peterson. "Early Parent-Child Interactions and Early Literacy Development." *Topics in Early Childhood Special Education*, vol. 23, no. 3, 2003, pp. 124–136.

Drehobl, K. F., and M. G. Fuhr. *Pediatric Massage for the Child with Special Needs.* Austin, TX: Pro-Ed, 2000

Feldman, R., A. I. Eidelman, and N. Rotenberg. "Parenting Stress, Infant Emotional Regulation, Maternal Sensitivity, and the Cognitive Development of Triplets: A Model for Parent and Child Influences in a Unique Ecology. *Child Development*, vol. 75, no. 6, 2004, pp. 1774–1791.

Feldman, R., A. I. Eidelman, L. Sirota, and A. Weller. "Comparison of Skin-to-Skin (Kangaroo) and Traditional Care: Parenting Outcomes and Preterm Infant Development." *Pediatrics*, vol. 110, 2002, pp. 16–26.

Fenson, L., and D. S. Ramsey. "Decentration and Integration of the Child's Play in the Second Year." *Child Development*, vol. 51, 1980, pp. 171–178.

Field, T. *Touch*. Boston, MA: Massachusetts Institute of Technology, 2001.

Field, T., D. Lasko, P. Mundy, T. Henteleff, S. Kabat, S. Talpins, and M. Dowling. "Brief Report: Autistic Children's Attentiveness and Responsivity Improve After Touch Therapy." *Journal of Autism & Developmental Disorders,* vol. 27, no. 3, 1997, pp. 333-8.

Field, T., N. Grizzle, F. Scafidi, S. Abrams, S. Richardson, C. Kuhn, and S. Schanberg. "Massage Therapy for Infants of Depressed Mothers." *Infant Behavior and Development,* vol. 19, no. 1, 1996, pp. 107-112.

Forbes, E. E., J. F. Cohn, N. B. Allen, and P. M. Lewinsohn. "Infant Affect During Parent-Infant Interaction at 3 and 6 months: Differences Between Mothers and Fathers and Influence of Parent History of Depression." *Infancy,* vol. 5, no. 1, 2004, pp. 61–84.

Forry, N.D., I. Iruka, K. Kainz, K. Tout, J. Torquati, A. Susman-Stilman, D. Bryant, R. Starr, and S. Smith. "Identifying Profiles of Quality in Home-Based Child Care." Issue Brief OPRE 2012-20. Washington, DC: Office of Planning, Research and Evaluation in the Administration for Children and Families, U.S. Department of Health and Human Services, 2012.

Forry, N.D., S. Moodie, S. Simkin, and L. Rothenberg. "Family-Provider Relationships: A Multidisciplinary Review of High Quality Practices and Associations with Family, Child, and Provider Outcomes." Issue Brief OPRE 2011-26a. Washington, DC: Office of Planning, Research and Evaluation in the Administration for Children and Families, U.S. Department of Health and Human Services, 2011.

Fuligni, A. S., W. J. Han, and J. Brooks-Gunn. "The Infant-Toddler HOME in the 2nd and 3rd Years of Life." *Parenting: Science and Practice,* vol. 4, no. 2–3, 2004, pp. 139–159.

Fuller, B., S. L. Kagan, S. Loeb, and Y-W Chang. "Child Care Quality: Centers and Home Settings that Serve Poor Families." *Early Childhood Research Quarterly,* vol. 19, no. 4, 2004, pp. 505–527.

Gelman, R. "First Principles Organize Attention to and Learning About Relevant Data: Number and Animate-Inanimate Distinction as Examples." *Cognitive Science,* vol. 14, 1990, pp. 79–106.

Grehan, Anna, and Allan Sterbinsky. Literacy Observation Tool Reliability Study. Paper presented at the annual meeting of the American Educational Research Association, Montreal, Canada, 2005.

Guralnick, M. J., B. Neville, R. T. Connor, and M. A. Hammond. "Family Factors Associated with Peer Social Competence of Young Children with Mild Delays." *American Journal on Mental Retardation*, vol. 108, no. 4, 2003, pp. 272–287.

Halle, T., J. E. Vick Whittaker, and R. Anderson. "Quality in Early Childhood Care and Education Settings: A Compendium of Measures, Second Edition." Washington, DC: Child Trends, 2010.

Halle, T., R. Anderson, A. Blasberg, A. Chrisler, and S. Simkin. "Quality of Caregiver-Child Interactions for Infants and Toddlers (Q-CCIIT): A Review of the Literature." OPRE 2011-25. Washington, DC: Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services, 2011.

Harms, T., D. Cryer, and R.M. Clifford. *Infant/Toddler Environment Rating Scale - Revised Edition*. New York, NY: Teachers College Press, 2003.

Harms, T., D. Cryer, and R.M. Clifford. *Infant/Toddler Environment Rating Scale - Revised Edition, Updated*. New York, NY: Teachers College Press, 2006.

Harms, T., D. Cryer, and R.M. Clifford. *Family Child Care Environment Rating Scale – Revised Edition*. New York, NY: Teachers College Press, 2007.

Hart B., and T. R. Risley. *Meaningful Differences in the Everyday Experience of Young American Children*. Baltimore, MD: Paul H. Brooks Publishing, 1995.

Heath, S. B. "What No Bedtime Story Means: Narrative Skills at Home and School." In *Language Socialization Across Cultures. Studies in Social and Cultural Foundation of Language,* edited by B. B. Schieffelin and E. Ochs. New York, NY: Cambridge University Press, 1986.

Hernandez-Reif, M., T. Field, S. Largie, S. Hart, M. Redzepi, B. Nierenberg, M. Peck. "Psychosocial Forum. Childrens' Distress During Burn Treatment is Reduced by Massage Therapy." *Journal of Burn Care & Rehabilitation,* vol. 22, no. 2, 2001, pp. 191-5.

Hofer, Kerry G. "How Measurement Characteristics Can Affect ECERS-R Scores and Program Funding." *Contemporary Issues in Early Childhood,* vol. 11, no. 2, 2010, pp. 175–191.

Howes, C. "Patterns of Friendship." *Child Development*, vol. 54, no. 4, 1983, pp. 1041–1053.

Howes, C. "Peer Interaction of Young Children." *Monographs of the Society for Research in Child Development*, Serial No. 217, vol. 53, no. 1, 1988.

Howes, C. "Peer Play Scale as an Index of Complexity of Peer Interaction." *Developmental Psychology*, vol. 16, no. 4, 1980, pp. 371–372.

Howes, C., and C. C. Matheson. "Sequences in the Development of Competent Play with Peers: Social and Social Pretend Play." *Developmental Psychology*, vol. 28, no. 5, 1992, pp. 961–974.

Howes, C., O. Unger, and L. Beizer Seidner. "Social Pretend Play in Toddlers: Parallels with Social Play and with Solitary Pretend." *Child Development*, vol. 60, no. 1, 1989, pp. 77–84.

Hudson, J. A. "The Emergence of Autobiographical Memory in Mother-Child Conversation." In *Knowing and Remembering in Young Children.* Edited by R. Fivush and J. A. Hudson. Cambridge, UK: Cambridge University Press, 1990.

Hurtado, N., V. A. Marchman, and A. Fernald. "Does Input Influence Uptake? Links Between Maternal Talk, Processing Speed and Vocabulary Size in Spanish-Learning Children." *Developmental Science*, vol. 11, no. 6, 2008, pp. F31–F39.

Huttenlocher, J., M. Vasilyeva, and E. Cymerman. "Language Input and Child Syntax." *Cognitive Psychology*, vol. 45, 2002, pp. 337–374.

Huttenlocher, J., W. Haight, A. Bryk, M. Seltzer, and T. Lyons. "Early Vocabulary Growth: Relation to Language Input and Gender. *Developmental Psychology*, vol. 27, no. 2, 1991, pp. 236–248.

Ispa, J. M., M. A. Fine, L. C. Halgunseth, S. Harper, J. Robinson, L., Boyce, J. Brooks-Gunn, and C. Brady-Smith. "Maternal Intrusiveness, Maternal Warmth, and Mother-Toddler Relationship Outcomes: Variations Across Low-Income Ethnic and Acculturation Groups." *Child Development*, vol. 75, no. 6, 2004, pp. 1613–1631.

Johnston, J. R., and M. A. Wong. "Cultural Differences in Beliefs and Practices Concerning Talk to Children." *Journal of Speech, Language & Hearing Research*, vol. 45, no. 5, 2002, pp. 916–926.

Joseph, G., P. Strain, and M. M. Ostrosky. "Fostering Emotional Literacy in Young Children: Labeling Emotions." CSEFEL What Works Brief #21. Available online at [http://csefel.vanderbilt.edu/briefs/wwb21.pdf], Accessed September 23, 2011.

Kaler, S. R., and B. J. Freman. "Analysis of Environmental Deprivation: Cognitive and Social Development in Romanian Orphans." *Journal of Child Psychology and Psychiatry,* vol. 35, no. 4, 1994, pp. 769-781

Kane, Thomas, J. and Douglas O. Staiger. "Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains." Seattle, WA: Bill and Melinda Gates Foundation, January 2012.

Kelly, J. F., and K. E. Barnard. "Assessment of Parent-Child Interaction: Implications for Early Intervention." In *Handbook of Early Childhood Intervention*, edited by J. P. Shonkoff and S. Meisels. Cambridge, UK: Cambridge University Press, 2000.

Kryzer, Erin M., N. Kovan, D. A. Phillips, L. A. Domagall, and M. R. Gunnar. "Toddlers' and Preschoolers' Experience in Family Day Care: Age Differences and Behavioral Correlates." *Early Childhood Research Quarterly*, vol. 22, no. 4, 2007, pp. 451–466.

Ladd, G. W. "Themes and Theories: Perspectives on Processes in Family-Peer Relationships." In *Family-Peer Relationship: Modes of Linkage*, edited by R. A. Parke and G. W. Ladd. Hillsdale, NJ: Erlbaum Associates, 1992.

Ladd, G. W., and C. H. Hart. "Creating Informal Play Opportunities: Are Parents' and Preschoolers' Initiations Related to Children's Competence with Peers?" *Developmental Psychology*, vol. 28, 1992, pp. 1179–1187.

Ladd, G. W., C. H. Hart, E. M. Wadsworth, and B. S. Goiter. "Preschoolers' Peer Networks in Nonschool Settings: Relationship to Family Characteristics and School Adjustment." In *Social Networks of Children, Adolescents, and College Students*, edited by S. Salzinger and J. Antrobus. Hillsdale, NJ: Erlbaum Associates, 1988.

Landry, S. H. "Preterm Infants Responses in Joint Attention Interactions." *Infant Behavior and Development*, vol. 9, 1986, pp. 1–14.

Landry, Susan H., K. E. Smith; C. L. Miller-Loncar; P. R. Swank. Predicting cognitive-language and social growth curves from early maternal behaviors in children at varying degrees of biological risk. *Developmental Psychology* ,1997 vol. 33, no.6, 1997, pp. 1040-53.

Linacre, J.M. "Sample Size and Item Calibration Stability." *Rasch Measurement Transactions*, vol. 7, no. 4, 1994, p. 328.

Linebarger, Deborah L., and Dale Walker. "Infants' and Toddlers' Television Viewing and Language Outcomes." *American Behavioral Scientist*, vol. 48, no. 5, 2005, pp. 624–645.

Lobo, M. A., and J. C. Galloway. "Postural and Object-Oriented Experiences Advance Early Reaching, Object Exploration, and Means-End Behavior." *Child Development*, vol. 79, no. 6, 2008, pp. 1869–1890.

Lonigan, C. J. "Reading to Preschoolers Exposed: Is the Emperor Really Naked?" *Developmental Review*, vol. 14, 1994, pp. 303–323.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. New Jersey: Addison-Wesley.

Mahoney, G., Spiker, D., & Boyce, G. (1996). Clinical assessments of parent-child interaction: Are professionals ready to implement this practice? *Topics in Early Childhood Special Education, v*ol. 16, no. 1, 1996, pp. 26-50.

Marfo, K., C. F. Dedrick, and N. Barbour. "Mother Children Interactions and the Development of Children with Mental Retardation." In *Handbook of Mental Retardation and Development*, edited by J. A. Burack, R. M. Hodapp, and E. Zigler. Cambridge, UK: Cambridge University Press, 1998.

Markus, J., P. Mundy, M. Morales, C. E. F. Delgardo, and M. Yale. "Individual Differences in Infant Skills as Predictors of Child-Caregiver Joint Attention and Language." *Social Development*, vol. 9, no. 3, 2000, pp. 302–315.

Martin, A., R. M. Ryan, and J. Brooks-Gunn. "The Joint Influence of Mother and Father Parenting on Child Cognitive Outcomes at Age 5." *Early Childhood Research Quarterly*, vol. 22, 2007, pp. 423–439.

Meisels, Samuel J., Sally Atkins-Burnett, and Julie Nicholson. *Assessment of Social Competence, Adaptive Behaviors, and Approaches to Learning with Young Children*. US Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics, 1996.

Montagu, A. *Touching: The Human Significance of the Skin*. New York: Harper & Row Publishers, 1971.

Mulligan, G.M., D. Brimhall, and J. West. "Child Care and Early Education Arrangements of Infants Toddlers, and Preschoolers: 2001 (NCES 2006-039)." U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office, 2005.

Munson, Leslie J., and Samuel L. Odom. "Review of Rating Scales that Measure Parent-Infant Interaction." *Topics in Early Childhood Special Education,* vol. 16, no. 1, 1996, pp. 1–25.

Muthén, L.K., and B.O. Muthén. *Mplus User's Guide, 5th ed*. Los Angeles: Muthén & Muthén, 2007.

National Research Council. *Working Families and Growing Kids: Caring for Children and Adolescents.* Committee on Family and Work Policies, E. Smolensky and J. Goodman (editors), Board on Children, Youth and Families, Division of Behavioral Sciences and Education. Washington, DC: The National Academies Press, 2003.

National Scientific Council on the Developing Child. **"**Children's Emotional Development Is Built into the Architecture of Their Brains." Working Paper No. 2**.** Cambridge, MA: National Scientific Council on the Developing Child, Center on the Developing Child at Harvard University, 2004b.

National Scientific Council on the Developing Child. **"**Establishing a Level Foundation for Life: Mental Health Begins in Early Childhood." Working Paper No. 6**.** Cambridge, MA: National Scientific Council on the Developing Child, Center on the Developing Child at Harvard University, 2012.

National Scientific Council on the Developing Child. "The Timing and Quality of Early Experiences Combine to Shape Brain Architecture." Working Paper No. **5**. Cambridge, MA: National Scientific Council on the Developing Child, Center on the Developing Child at Harvard University, 2007.

National Scientific Council on the Developing Child. "Young Children Develop in an Environment of Relationships." Working Paper No. 1**.** Cambridge, MA: National Scientific Council on the Developing Child, Center on the Developing Child at Harvard University, 2004.

NICHD Early Child Care Research Network. "Characteristics of Infant Child Care: Factors Contributing to Positive Caregiving." *Early Childhood Research Quarterly*, vol. 11, 1996, pp. 269–306.

NICHD Early Child Care Research Network. "Child-Care Structure -> Process ->Outcome: Direct and Indirect Effects of Child-Care Quality on Young Children's Development." *Psychological Science*, vol. 13, no. 3, 2002a, pp. 199–206.

NICHD Early Child Care Research Network. "Early Child Care and Children's Development Prior to School Entry: Results from the NICHD Study of Early Child Care and Youth Development." *American Educational Research Journal*, vol. 39, no. 1, 2002b, pp. 133–164.

NICHD Early Child Care Research Network. "The Relation of Child Care to Cognitive and Language Development." *Child Development*, vol. 71, 2000, pp. 958–978.

Nunnally, J., & Bernstein, I. (1994). *Psychometric Theory* (3rd ed., p. 736). McGraw-Hill Humanities/Social Sciences/Languages.

O'Connell, B., and I. Bretherton, I. "Toddlers' Play, Alone and with Mother: The Role of Material Guidance." In *Symbolic Play: The Development of Social Understanding* edited by I. Bretherton. New York: Academic Press, 1984.

O'Reilly, A. W., and M. H. Bornstein. "Caregiver-Child Interaction in Play." *New Directions for Child Development*, vol. 59, 1993, pp. 55–66.

Ostrosky, M. M., E. Y. Jung, M. L. Hemmeter, and D. Thomas. "Helping Children Understand Routines and Classroom Schedules" CSEFEL What Works Brief #3. Available online at http://csefel.vanderbilt.edu/briefs/wwb3.pdf, Accessed September 23, 2011a.

Ostrosky, M. M., M. L. Hemmeter, J. Murry, and G. Cheatham. **"**Helping Children Express Their Wants and Needs." CSEFEL What Works Brief #19. Available online at http://csefel.vanderbilt.edu/briefs/wwb19.pdf, Accessed September 23, 2011b.

Parke, R. A., S. M. Profilet, and G. W. Ladd. "Parents' Management of Children's Peer Relations: Facilitating and Supervising Children's Activities in the Peer Culture." In *Family-Peer Relationship: Modes of Linkage,* edited by R. A. Parke and G. W. Ladd. Hillsdale, NJ: Erlbaum Associates, 1992.

Patterson, H. D. and Thompson, R. (1971), "Recovery of Inter-Block Information When Block Sizes Are Unequal," *Biometrika*, 58, 545–554.

Perry, B. D., and D. Pollard. *Altered Brain Development Following Global Neglect in Early Childhood.* Society for Neuroscience: Proceedings from Annual Meeting, New Orleans, 1997.

Perry, B., and M. Szalavitz. *The Boy Who Was Raised as a Dog: And Other Stories from a Child Psychiatrist's Notebook – What Traumatized Children Can Teach Us About Loss, Love and Healing*. New York, NY: Basic Books, 2007.

Phillipsen, L. C., M. R. Burchinal, C. Howes, and D. Cryer. "The Prediction of Process Quality from Structural Features of Child Care." *Early Childhood Research Quarterly,* vol. 12, no. 3, 1997, pp. 281–303.

Poehlmann, J., and B. H. Fiese. "Parent-Infant Interaction as a Mediator of the Relation Between Neonatal Risk Status and 12-month Cognitive Development." *Infant Behavior and Development*, vol. 24, 2001, pp. 171–188.

Ratner, H. H. "Memory Demands and the Development of Young Children's Memory." *Child Development*, vol. 55, 1984, pp. 2173–2191.

Rathmann, Peggy. *Good Night, Gorilla.* New York: Putnam Juvenile, 1996.

Raudenbush, S.W., and S. Sadoff. "Statistical Inference when Classroom Quality is Measured with Error." *Journal of Research on Educational Effectiveness*, vol. 1, 2008, pp. 138–154.

Ruff, H. A. "Attention and Organization of Behavior in High-Risk Infants." *Journal of Developmental and Behavioral Pediatrics*, vol. 7, 1986, pp. 298–301.

Ryan, R. M., A. Martin, and J. Brooks-Gunn. "Is One Good Parent Good Enough? Patterns of Mother and Father Parenting and Child Cognitive Outcomes at 24 and 36 Months." *Parenting: Science and Practice*, vol. 6, no. 2–3, 2006, pp. 211–228.

Sabol, T. J., SL Soliday Hong, R. C. Pianta, and M. R. Burchinal. "Can Rating Pre-K Programs Predict Children's Learning?" *Science,* vol. 341, no. 6148, 2013, pp. 845–846.

Sameroff, A. J., and M. J. Chandler. "Reproductive Risk and the Continuum of Caretaking Casualty." In *Review of Child Development Research*, vol. 4, edited by F. D. Horowitz, M. Hetherington, S. Scarr-Salapatek, and G. Sigel. Chicago, IL: University of Chicago Press, 1975.

Sameroff, A. *The Transactional Model of Development: How Children and Contexts Shape Each Other*. Washington, DC: American Psychological Association, 2009.

Schneider, E. F., *Massaging your Baby: The Joy of Touch Time.* Garden City Park, NY: Square One Publishers, 2006.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability Theory: A Primer*. Newbury Park, CA: Sage.

Shonkoff, J. P., and D. A. Phillips. *From Neurons to Neighborhoods*. Washington, DC: National Academies Press, 2000.

Shonkoff, Jack P. "From Neurons to Neighborhoods: Old and New Challenges for Developmental and Behavioral Pediatrics." *Journal of Developmental & Behavioral Pediatrics*, vol. 24, no. 1, 2003, pp. 70–77.

Sim, Julius, and Chris C. Wright. "The Kappa Statistic in Reliability Studies: Use, Interpretation, and Sample Size Requirements." *Physical Therapy*, vol. 85, no. 3, 2005, pp. 257–268.

Snow, C. E., M. S. Burns, and P. Griffin. *Preventing Reading Difficulties in Young Children*. Washington, DC: National Academy Press, 1998.

Steelman, L. M., M. A. Assel, P. R. Swank, K. E. Smith, and S. H. Landry. "Early Maternal Warm Responsiveness as a Predictor of Child Social Skills: Direct and Indirect Paths of Influence Over Time." *Applied Developmental Psychology*, vol. 23, 2002, pp. 135–156.

Tamis-LeMonda, C. S., and M. H. Bornstein. "Individual Variation, Correspondence, Stability, and Change in Mother-Toddler Play." *Infant Behavior and Development*, vol. 14, 1991, pp. 143–162.

Tamis-LeMonda, C. S., J. D. Shannon, N. J. Cabrera, and M. E. Lamb. "Fathers and Mothers at Play with Their 2- and 3-Year-Olds: Contributions to Language and Cognitive Development." *Child Development*, vol. 75, no. 6, 2004, pp. 1806–1820.

Tamis-LeMonda, C. S., M. H. Bornstein, and L. Baumwell. "Maternal Responsiveness and Children's Achievement of Language Milestones." *Child Development*, vol. 72, no. 3, 2001, pp. 748–767.

Tomopoulos, S., B.P. Dreyer, S. Berkule, A.H. Fierman, C. Brockmeyer, A.L. Mendelsohn. "Infant Media Exposure and Toddler Development." *Archives of Pediatric and Adolescent Medicine*, vol. 164, 2010, pp. 1105–1111.

Vandell, D. L., and B. Wolfe. "Child Care Quality: Does It Matter and Does It Need to Be Improved?" Madison, WI: University of Wisconsin-Madison, Institute for Research on Poverty, 2000.

Wang, Yan Z., Angela R. Wiley, and Xiaobin Zhou. "The Effect of Different Cultural Lenses on Reliability and Validity in Observational Data: The Example of Chinese Immigrant Parent–Toddler Dinner Interactions." *Social Development*, vol. 16, no. 4, 2007, pp. 777–799.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2007). Reliability Coefficients and Generalizability Theory. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics*, Vol. 26 (pp. 81-124).

Winsteps. "Table 30.1 Differential Item Functioning DIF Pairwise." Available at http://www.winsteps.com/winman/table30_1.htm.

Winton, P. J., J. A. McCollum, and C. Catlett. *Practical Approaches to Early Childhood Professional Development: Evidence, Strategies, and Resources*. Washington, DC: Zero to Three, 2008.

Zaslow, M., I. Martinez-Beck, K. Tout, and T. Halle. *Quality Measurement in Early Childhood Settings*. Baltimore, MD: Brookes, 2011.

Zeanah, C. H., O. Mammen, and A. Lieberman. "Disorders of Attachment." In *Handbook of Infant Mental Health*, edited by S. Zeanah. New York, NY: Guilford Press, 1993.

Zwick, R., D.T. Thayer, and C. Lewis. "An Empirical Bayes Approach to Mantel-Haenszel DIF Analysis." *Journal of Educational Measurement,* vol. 36, no. 1, 1999, pp. 1–28.

This page has been left blank for double-sided copying.


Q-CCIIT