Infant–toddler teachers can successfully employ authentic assessment: The Learning Through Relating system

Amanda J. Moreno, Mary M. Klute

Abstract

This study documents the reliability and validity of a new infant–toddler authentic assessment, the Learning Through Relating Child Assets Record (LTR-CAR), and its feasibility of use by infant–toddler caregivers in an Early Head Start program. In a sample of 136 children, results indicated a strong internal structure of the LTR-CAR as evidenced by high internal consistency within age bands, expected performance in terms of the order of difficulty of items, and expected performance relative to the chronological ages of children. Moreover, the LTR-CAR accounted for a statistically significant amount of unique variance (between 6 and 26%), over and above age and gender, in six out of eight regressions predicting criterion measures. Given the increasing numbers of infants and toddlers in non-parental care, the need for authentic assessment as a tool to plan meaningful learning opportunities, as well as the lack of evidence-based authentic assessment options at the infant–toddler level, the LTR-CAR may provide a viable option for the field.

© 2011 Published by Elsevier Inc.

According to the National Association for the Education of Young Children’s position statement on developmentally appropriate practice (Copple & Bredekamp, 2009), “Teachers cannot be intentional about helping children to progress unless they know where each child is with respect to learning goals” (p. 22). Despite the recognition that individualized assessment of learning and appropriate curricular planning go hand in hand, reliable, valid, and authentic options for implementing this cyclical system are extremely limited, particularly at the infant–toddler level (Gilbert, 2001). This is puzzling given the fact that decades-old research indicates that this age-range is the highest in vulnerability and potential, and also the most experientially determined period of brain development in the human life span (e.g., Greenough, Black, & Wallace, 1987; Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991; Kuhl, 1993). Although the majority of all children under three in the United States are regularly in the care of individuals who are not their parents (Kim & Peterson, 2008), cultural ambivalence still remains as to why such caregiving experiences should be any more than “instinctive” parenting, which should provide evolutionarily guaranteed and at least adequate stimulation. In other words, for typically developing children in the care of non-parents, one might wonder why a sensitive and warm adult is not sufficient, and why planning based on assessment is necessary at this early age.

Although good teaching and good parenting do share many features across ages of children (Moore, 2006), these learning contexts are far from identical from either the child’s or the adult’s perspective. These obvious differences (e.g., someone else’s children vs. one’s own, group-based vs. one-on-one interactions, part of the day vs. all of it), coupled with evidence that average non-parental early learning environments are of moderate or poor quality (e.g., LoCasale-Crouch et al., 2007; Peisner-Feinberg & Burchinal, 1997; Whitebook, Howes, & Phillips, 1990), especially for infant–toddler settings (Helburn et al., 1995; WestEd, 2002), indicate that there is a clear need to implement strategies that do not rely solely on the good instincts of infant–toddler teachers. Contrary to a common belief, intentional instruction need not imply highly prescriptive adult-centered regimens, and learning assessments do not necessitate inappropriate “testing” of infants and toddlers (Tzuo, 2007). This holds true to the extent that infant–toddler “instruction” is not merely a downward extension of preschool (Lally, 2009; Raikes & Edwards, 2009) and that assessments are developmentally appropriate and undertaken for the purpose of increasing the degree of reflection and intentionality in teacher–child interactions (Bowers, 2008; Mogharreban, McIntyre, & Raison, 2010). In short, high-quality infant–toddler learning environments require reflexive, individualized planning of appropriate learning experiences.
via the systematic observation of children’s strengths and needs (Bagnato, Neisworth, & Munson, 1997; Bowers, 2008; Grisham-Brown, Hallam, & Brookshire, 2006; Meisels, 1999).

Tools and methods are needed in order to carry out credible observations as a part of best practice, but programs struggle to find assessment systems that meet all of their needs (e.g., developmentally appropriate, valid and reliable, supportive of family-program connections, aligned with curriculum, learning standards, and accountability initiatives, etc.). Moreover, the limited arsenal of tools designed for implementation by non-specialized infant–toddler caregivers lacks an evidence base comparable in breadth and quality to that which exists for tools available to preschool teachers (e.g., The Work Sampling System, Meisels, Jablon, Marsden, Dichtelmiller, & Dorfman, 2001) or infant–toddler interventionists working in early intervention programs (e.g., Assessment, Evaluation, and Programming System for Infants and Toddlers, Bricker, 2002). Thus, the Learning through Relating Child Assets Record (LTR-CAR; Moreno, Sciarra, & Klute, 2009) was developed to provide a viable option for non-specialized infant–toddler caregivers to authentically assess children in their care with or without developmental delays.

The current study presents findings from two years of implementation of the LTR-CAR. The research questions driving this study are about the psychometric properties of the scores obtained from the LTR-CAR checklist, and its feasibility for use by infant–toddler teachers and home visitors. Our focus in this manuscript is not on the instrument itself needed to be examined before its usage as a planning tool could be explored. It is important to first assess whether infant–toddler caregivers with limited formal education and no specialized early intervention experience can make accurate assessments about the children in their charge using the LTR-CAR.

1. Authentic assessment for infants and toddlers

Meisels (1998) provides a concise definition of authentic assessment in early childhood, which is the careful observation of children’s skills “in situ [in their proper place], over time, and differentially” (p. 21). Bagnato (2007) further posits that authentic assessment is conducted by “… familiar and knowledgeable caregivers about the naturally occurring competencies of young children in daily routines” (p. 27). Thus, in addition to the authentic nature of test items and methodologies, a relational context of assessment is key to true authenticity. For our part, we are interested in the point on the continuum of authenticity at which the assessors are individuals whose primary purpose is to regularly and frequently interact with the child in the natural settings in which development unfolds (e.g., a parent, a child care provider). Although the majority of work on infant–toddler authentic assessment comes out of the tradition of early childhood intervention (e.g., Bagnato, Neisworth, & Pretti-Frontczak, 2010; Linder & Linas, 2009), the benefits of applying authentic methodologies to early educational settings more generally – for example, those that exist to provide child care for working families – are now also recognized (Dupnuy, 2008; Jablon, Dombro, & Dichtelmiller, 1999; Kline, 2008; Meisels, Wen, & Beachy-Quick, 2010). Since infants and toddlers are now frequently in the care of non-parents either for child care purposes or because of family-related risk factors (e.g., Early Head Start), and not necessarily because they have been referred for developmental concerns, it is important to know whether the caregivers in these settings, who often have less training than early interventionists (Bagnato & Ethridge, 2010; Hestenes, Cassidy, Hedge, & Lower, 2007), can effectively employ authentic assessment.

The merits of authentic assessment over standardized assessment for children prior to school entry have been documented extensively elsewhere (e.g., Bagnato, 2007; Grisham-Brown et al., 2006; Kline, 2008; Macy & Bagnato, 2010; Meisels, 1996, 2007; Neisworth & Bagnato, 2004; Sandall, Hemmeter, Smith, & McLean, 2005). Due to a multitude of drawbacks of standardized cognitive and developmental testing, including its limited utility and predictive validity for preschoolers and kindergarteners (Kim & Suen, 2003; LaParo & Pianta, 2000), the inadequate involvement of families in such procedures, and the promotion of “non-functional” skills with no real-world value (Neisworth & Bagnato, 2004), authentic assessment is now the recommended and accepted practice (Copple & Bredekamp, 2009; Grisham-Brown, Hallam, & Pretti-Frontczak, 2008; Meisels et al., 2010; National Association for the Education of Young Children & National Association of Early Childhood Specialists in State Departments of Education, 2003). Beyond being the preferred format of assessment, authentic methodology is now considered to be central to providing high-quality educational experiences in early childhood because it is “connected to specific, beneficial purposes [such as] making sound decisions about teaching and learning” (National Association for the Education of Young Children & National Association of Early Childhood Specialists in State Departments of Education, 2003; p. 2).

2. Accountability and the evidence base

In addition to the benefits to children provided by authentic assessment, many initiatives are seeking to use it as part of increased efforts to document accountability (Meisels, 2007). Although authentic developmental assessments are not commonly used for high-stakes purposes such as granting or denying services to individual children, their use is still going strong for broader purposes such as the evaluation of preschool programs, competitions for limited early learning funding, and statewide progress monitoring (e.g., the Results Matter initiative in Colorado). This is cause for concern considering that few of these assessments have peer-reviewed evidence documenting that they can be reliably and validly employed by the broad swath of caregivers and teachers expected to implement them.

Very few well-known, teacher-employed, authentic assessments have such evidentiary support. For example, The Child Observation Record (COR; High/Scope Educational Research Foundation, 1992) and the Work Sampling System (WSS; Meisels et al., 2001) have published validation studies. However, both of these authentic assessments are appropriate for children in the preschool age range and do not fill the need for caregivers of infants and toddlers. The High/Scope COR has an infant–toddler version but published evidence of its psychometric properties is limited to one non-refereed study available only as an appendix on the High/Scope website (High/Scope Educational Research Foundation, 2002). The study used a sample of 50 infants and toddlers for reliability, and a subset of 30 for validity. Reliability coefficients were over .90, and after removing the effects of age, the COR was significantly correlated at .36 with Bayley Scales of Infant Development (Bayley, 1993) mental age but not significantly correlated with Bayley motor age.

Another widely used authentic assessment endorsed by many statewide accountability initiatives, The Creative Curriculum Developmental Continuum (CCDC; Teaching Strategies, 2002), is available in both preschool and infant–toddler versions but neither has any peer-reviewed evidence documenting reliability or validity. The newest version, which combines the infant–toddler and preschool age ranges (i.e., Teaching Strategies GOLD) reports...
an ongoing field test on more than 2000 children on its website (http://www.teachingstrategies.com/page/GOLDFAQ.cfm#research2) but it is not stated whether this test includes an external validity component, or when initial results might be expected.

The Ounce Scale (Meisels, Dombro, Marsden, Weston, & Jewkes, 2003) was designed specifically for infants and toddlers (birth to age 3 1/2), although evidence supporting it was not published until it had been in wide use for seven years (Meisels et al., 2010). This recently published study presents the best available evidence on the reliability and validity of an authentic assessment for infants and toddlers (not specific to early intervention settings) given that it was peer-reviewed and employed a substantial sample size (n = 265). Average reliability across items and within age band, as indicated by Cronbach’s alpha, proved to be low: the mean alpha was .56 (range .19–.89). However, the reliability may have suffered because the analyses were limited by the smaller sub-samples in each age band (n’s from 26 to 38), given the structure of the Ounce Scale in separate assessment “booklets” by age. In other words, not all children are assessed on all items. The validity evidence for the Ounce Scale was stronger. Bivariate correlations between the Ounce Developmental Profile and the criterion measures including the Bayley Mental and Physical Developmental Indexes (Bayley, 1993), Preschool Language Scales, 4th Edition (PLS-4; Zimmerman et al., 1992), and the Ages and Stages Questionnaire – Social Emotional (Squires, Bricker, & Twombly, 2002) were in the expected direction, statistically significant, and ranged in size from .28 to .47. Moreover, the Ounce Scale and the criterion measures agreed in their classification of children as delayed vs. non-delayed between 73 and 76% of the time. However, in the hierarchical regressions predicting the criterion measures from the Ounce Scale, only betas and their significance were reported, and not percent of variance. Since chronological age subsumes a significant portion of variance, it is important to know how much of the unique variance was accounted for by the Ounce Scale.

3. Background and overview of Learning Through Relating

The LTR-CAR was designed for use by infant–toddler caregivers with or without college degrees. Consistent with the aim of authentic assessment, the LTR-CAR was developed as a mechanism for infant–toddler caregivers to: (a) learn in a systematic way about the children in their charge in terms of their strengths, style, preferences, and areas where they may need further support; (b) plan age-appropriate and child-centered learning opportunities for each child that flow from what is learned from the observation and documentation embedded in the tool; and (c) support parents and families in learning about their child’s development and the best ways to support it. Although the data reported here refer only to the authentic assessment checklist, it is important to note that the LTR-CAR is just one part of the larger LTR learning system. Overall, the LTR system includes comprehensive curriculum areas and activities designed for children, goals and professional development guidance for adult caregivers facilitated by ongoing, weekly coaching and reflective supervision, and an individualized planning tool (which includes recommended next steps for both children and their caregivers) that is linked systematically to the observations made in the CAR. It is designed to support caregivers’ engagement in meaningful interactions with children that in turn enhance children’s age-appropriate language skills, frequency and variety of book-engagement and pre-writing behaviors, and clarity and regulation of interactions with peers and adults.

The LTR-CAR is different from many developmental assessments, which typically attempt to cover discrete areas of development such as social-emotional, communication and language, cognitive development, and physical development. In contrast, and based on the work of several scholars including Campos and colleagues (e.g., Campos et al., 2000), Rose and colleagues (e.g., Rose, Feldman, & Jankowski, 2009), Gopnik (2009), and Mandler and colleagues (e.g., Mandler & McDonough, 1993), LTR assumes a domain-general approach to development in the first three years of life. That is, LTR aligns with the global nature of development and the “inborn curriculum” (Lally, 2009) of infants and toddlers, and is therefore concerned with the quality of learning, regardless of the specific domain or domains in which an activity might be categorized. Thus, rather than domains, LTR employs learning topics such as Play and Interacting in Tune. Play and attuned relationships are not domains of development, but rather generalized backdrops on which any type of learning can occur, and on which multiple types of learning regularly occur simultaneously. Because caregivers should learn to create non-compartmentalized learning experiences for infants and toddlers, which is an approach that is age-appropriate and supported by brain research (Bricker, 2002; Lally, 2009), the topics in LTR are simply a means of “chunking” (Miller, 1956) the curriculum for the adult learners, but are not intended to be orthogonal domains of child development. Therefore, the reliability and validity of scores obtained by teachers on the LTR-CAR was particularly important to study in light of its non-traditional approach.

Another aspect of the LTR system worth highlighting is the inclusion of features designed to reduce socially desirable responding on the part of the infant–toddler caregivers, specifically, the documented tendencies to: (a) score children according to their chronological age; (b) be unwilling to score the “not pass” or “not proficient” rating because they feel it will unfairly characterize children’s abilities; and (c) not see, or become inured to the struggles or delays of children with whom they have a close relationship (Meisels et al., 2010). First, chronological age is not mentioned or written anywhere on the LTR-CAR, manual, or training materials (other than that the system is designed for birth-to-three). Second, and more important than simply not mentioning age, the LTR-CAR assesses all children on all items. By way of comparison, the Ounce Scale has 12–16 items per age band; the LTR-CAR has only slightly more at 18 items per age band. While children assessed using the Ounce Scale are scored only on the 12–16 items in their chronological age band, all children assessed using the LTR-CAR are scored on 18 items in each of 8 age bands, or 144 items altogether. This method shares some of the strengths of basal/ceiling scoring methodologies, and supports the use of the LTR-CAR as a part of a professional development process that helps teach infant–toddler caregivers about developmental trajectories, which they in turn can use to help parents understand what comes before and after particular skills and why. Moreover, LTR’s de-emphasis on chronological age encourages a “whole-child” approach and may make it more appropriate than other authentic instruments for use with children with developmental delays. Although the burden on caregivers is greater with 144 items on the checklist, the task is nonetheless reasonable given the authentic assessment methodology, along with the 3–4 month timeframe to fill out one CAR (depending on whether a program chooses to administer it three or four times per year).

Finally, the “proficient” and “not proficient” categories are named, Got It (GI) and Open Opportunity (OO), and there are specific training experiences around the selection of these terms. Caregivers are trained to select Got It based only on the saliency of a particular behavior in the child’s repertoire, and not on its frequency. For example, caregivers learn that the degree of self-initiation a child exhibits for a particular behavior is more important than the number of times a child shows the behavior. In addition to the guidance manual that provides specific descriptors of what each skill looks like with a Got It score, a series of self-reflective questions around
the definition of saliency (e.g., “Is this behavior comfortable or typical for the child? (GI) vs. does it require support, prompting, several tries, or great effort? (OO’)”) is provided to help caregivers score accurately. Similarly, training and coaching processes also emphasize that a score of Open Opportunity does not indicate a failure on the child’s part or that the child has never shown the behavior, but rather, that the caregiver would like some more time to plan learning opportunities for the child around the behavior.

4. The present study

This study adds to a very small body of research that examines the reliability and validity of authentic assessments specifically designed for children birth-to-three with or without delays. Different from prior research, this study presents data at more than one time point (fall and spring) on both the infant–toddler authentic assessment and the criterion measures being used to examine validity, so that developmental change captured by the instruments could be evaluated. Like the Ounce Scale validation study, this study also took place in the context of Early Head Start. This setting is appropriate because the LTR-CAR was designed to be applicable to children from low-income families, with or without disabilities, and employable by infant–toddler caregivers with or without college degrees. Importantly, this study employed the LTR-CAR with caregivers and children in both of the primary service options for Early Head Start programs: home-based and center-based, thus examining for the first time whether infant–toddler home visitors, who typically see the children in their care for significantly less time per week than center-based teachers, can also accurately employ authentic assessment. This aspect is important due to the fact that many infants and toddlers receive their child care in home-based settings. The LTR-CAR has both common and unique features with the few other existing infant–toddler authentic assessments, and may offer a viable alternative for either structural or evidentiary reasons.

The aim of this study was to evaluate the reliability and validity of the scores obtained from the LTR-CAR checklist as implemented by Early Head Start teachers and home visitors. The first goal was to assess the extent to which LTR-CAR scores are reliable. Reliability was assessed via internal consistency across items within age bands. Because of the nature of the LTR-CAR assessment, it is not feasible to examine inter-rater reliability. When completing the LTR-CAR, teachers are asked to make assessments of children based on the wealth of observations that they have made over a period of time. As such, an independent outside rater could not feasibly rate a child using the LTR-CAR. An alternative would be to have other caregivers who regularly interact with the child provide a second rating for the purposes of inter-rater reliability. This approach is also flawed because inter-rater reliability assumes two independent observations of the same target. It would be very difficult to make a persuasive case that two co-teachers, who collaborate on a daily basis to plan for and care for a group of children, can be viewed as independent raters (see Meisels et al., 2010 for a discussion of this issue). Other studies have made use of videotaped standardized activities designed to elicit particular behaviors as a means to enable a second rater to assess preschool-aged children using an authentic assessment (Grisham-Brown et al., 2008). This approach is not appropriate for the LTR-CAR because teachers are asked to make judgments as to whether each skill is a salient part of a child’s behavioral repertoire. During the infant–toddler developmental period, seeing a behavior once in response to a standardized activity is not sufficient for evaluating its saliency. This can only be gleaned based on ongoing, truly authentic, observation of the child.

The second goal of this study was to assess the internal validity of the LTR-CAR, by focusing on four specific questions: (a) Do the items of the LTR-CAR go in the developmental order proposed?; (b) Do average scores increase across the academic year?; (c) Do child scores near or slightly below their chronological age, given the higher than average risk exemplified by an Early Head Start population?; and (d) Do children score closer to age-expected performance by spring than they did in the fall? This pattern would be expected because all of the children were enrolled in an EHS program. Since most EHS families bear some socio-demographic risk, we hypothesized that the children may start out lower than their chronological age, but make gains over the course of the year. This question is relevant to the psychometric properties of the LTR-CAR in that it shows that the assessment is sensitive to the effects of the intervention. The third goal was to assess the external validity of the LTR-CAR by examining associations between the LTR-CAR and three criterion measures selected to correspond with three areas of child development that the LTR system is designed to impact: language skills, engagement with books and pre-writing skills, and regulation of interactions with peers and adults. We examined the associations between the criterion measures and the LTR-CAR in two ways. First, we analyzed the association between the continuous variables, after accounting for demographic characteristics. Second, we examined the extent to which the LTR-CAR and the criterion measures lead to similar conclusions about whether a child is at risk for developmental delay.

5. Method

5.1. Participants

There were 136 children with LTR-CAR data in the fall and 123 in the spring. Primarily due to aging out of Early Head Start or leaving the program, a total of 98 of these children had CAR data at both time points: A smaller subset of children and families (n = 89 for fall, n = 70 for spring) agreed to the research home visits during which the criterion assessments were conducted. Fifty-two percent of the children were female. At the fall time point, 21.4% of the children were 0–12 months old, 36.9% were 13–24 months old, 35.7% were 25–36 months old, and 6% were greater than 36 months old (mean age = 21.96 months, range = 1 month–38 months). The ethnic breakdown of the children was as follows: 11.4% White/Caucasian, 43.0% Black/African-American, 44.3% Hispanic/Latino, and 1.3% Native American. According to parent report, 57.1% of the children were most comfortable speaking English, 36.3% were most comfortable speaking Spanish, and 6.6% were most comfortable speaking another language. Parents’ highest level of educational attainment was as follows: 31.9% had not graduated high school, 26.4% graduated high school or had a GED, 37.4% had some college or post high school vocational training, and 4.3% had a college degree (Associate’s or Bachelor’s). Parents, who depending on their primary language, took either the Letter-Word Identification subtest of the Woodcock–Johnson III Tests of Achievement (Woodcock, McGrew, & Mather, 2001) or Bateria III Woodcock–Muñoz (Muñoz-Sandoval, Woodcock, McGrew, & Mather, 2005), had a mean standard score of 90.2 (SD = 14.63).

The study was conducted in an Early Head Start program that provides three service options to families: home-based, center-based, and a combination option, which entails two mornings per week at the EHS center and a home visit every other week. A total of 23 professional caregivers from this program, all of whom were female, participated across the two years of the study and administered the LTR-CAR. Three of these caregivers provided services to families in the home-based option; the other 20 were either in the center-based or combination option. According to self-report, two caregivers had less than two years of child care/family support experience, three had 3–5 years experience, six had 6–10...
years experience, and 11 had more than 10 years experience (one
teacher did not answer). All but one caregiver had at least a high
school degree or GED. Four caregivers had some college, two had
a Child Development Associate (CDA) certificate, five had a two-
year college degree, three had a four-year college degree, and
five had at least some formal post-college education. Four care-
givers reported being White/Caucasian, six were Hispanic/Latino,
five were Black/African-American, one was Asian-Pacific Islander,
one was mixed/other, and six did not answer.

5.2. Procedures

The data reported here were drawn from a larger development
and evaluation study of the overall LTR learning system. That study
used a phased-in, quasi-experimental design, such that only half
of the EHS program employed LTR the first year, and the entire
program used it the second year. This is the reason why several
children were assessed with the criterion measures, but not the
LTR-CAR during the first year (since using the authentic assess-
ment was embedded in learning how to use the curriculum), and
therefore why the pair-wise sample size for the external validity
analyses is less than the full sample of children assessed with the
criterion measures.

Caregivers received two days of training on the LTR learning
system; one of those days was dedicated exclusively to authentic
assessment philosophy, definition, and methods; theoretical and
practical matters of observation and documentation (e.g., using
detailed description of child behaviors rather than assumptions
about the child’s feelings, methods of anecdote keeping such as
sticky notes, dry erase sheets); and the specific features of the LTR-
CAR. The other half of training focused on the curriculum elements
of the LTR system (e.g., planning scaffolded activities). This training
was designed to give caregivers an initial understanding of the LTR
system so they could begin to implement it in their work. How-
ever, the expectation was that ongoing support would be needed
for teachers to fully implement the LTR system. As a result, the
LTR system includes ongoing individual and team-based coaching.
This involved an hour per week of out-of-class (or home) reflective
supervision with a coach, and 2h per month of team-based coach-
ing called Booster Sessions. During both types of coaching, time
was set aside to discuss the proper use of the LTR-CAR, as well as
talk about individual children’s activities and progress, and how to
accurately score them on the skill-waves. Although caregivers sub-
mitted the LTR-CAR three times per year (fall, winter, and spring),
only the fall and spring data are reported here because these time-
points coincided with when the criterion measures were obtained.

All caregivers were taught to use discussions with parents as a
supplement to their own observations, but caregivers in the pro-
gram’s home-based service option had an increased need to rely
on parent report. To make use of parent feedback, caregivers were
coached not to ask a parent if a child does or does not do a spe-
cific item on the CAR, but rather to ask them to describe specific
examples of play episodes, e.g., “When your child reached for the
object, did he get it the first time or reach a few times first?
Did he smile when he got the object, or start doing something
with it, like mouthing it or banning it?” In addition, and consist-
tent with EHS Performance Standards, the home-based option also
included monthly socializations with other children and families
at the center. LTR-CAR items related to peer interactions were
assessed during socializations for exclusively home-based children.

Families were invited by research assistants to be contacted
about the research study. Those who gave their contact informa-
tion were called at a later date. It was explained to parents that the
study involved a home visit by two research assistants, once in the
fall and once again in the spring. If the family wanted to participate,
the first visit was scheduled, and formal informed consent proce-
dures were conducted on the first visit. Families were compensated
with grocery store gift certificates. The research home visit proto-
col involved a standardized child language assessment, a parent
interview, and a child–parent play and reading session. In addition,
for only one of the time points (most occurred during the fall), the
parent’s baseline literacy was also assessed during the visit.

Research assistants and coders were trained by the principal
investigator and a study coordinator with extensive experience
in coding and supervision of coding. Individuals were required to
videotape themselves administering the full protocol to at least five
pilot families and feedback was provided after each taping. One
measure involved behavioral coding from videotape. For this mea-
sure, coders were not allowed to code for the research study until
they achieved 80% agreement with the first author for three pilot
cases in a row after training.

5.3. Measures

Criterion measures. We included three criterion measures
relevant to three developmental areas addressed by the LTR-
CAR: language development, regulation of interactions with
peers and adults, and engagement with books and pre-writing
Zimmerman, Steiner, & Pond, 2002), the Competence Subscale of
the School–Toddler Social-Emotional Assessment (CS-ITSEA;
Carter & Briggs–Gowan, 2000), and a measure designed for this
study, the Assessment of Pre-Literacy Skills (APLS, The Apples;
Moreno, 2006). In all, our assessments of children for this study
included multiple reporters/viewpoints of the child including
objective experimenters (both standardized assessment and obser-
vational/behavioral coding), parents, and professional caregivers.

The PLS-4 is a direct, standardized assessment that our pro-
fessional research assistants administered in the child’s preferred
language. English or Spanish, in the family’s home. The fourth edi-
tion of this test was re-normed on over 1500 children including
39% ethnic minority children, and had high internal consistency,
test/re-test reliability, and accurate prediction of language delays
and disabilities (Zimmerman et al., 2002). PLS scores are reported
in both age-equivalent (in number of months) and standard scores
(with a standardized mean of 100 and standard deviation of 15).

We chose the CS-ITSEA because of its intuitive relationship to
the kinds of group-based learning readiness skills tapped by
the LTR-CAR such as child compliance, attention, imitation/play,
mastery motivation, empathy, and prosocial peer relations. It is scored
on a 0 (Not True/Rarely), 1 (Somewhat True/Sometimes), 2 (Very
True/Often) scale. The ITSEA was validated on a sample of over
1200 parent reports (Carter, Briggs–Gowan, Jones, & Little, 2003).
We administered the CS-ITSEA as an interview (asked the questions
aloud rather than gave a form to fill out) during the home visits, in
either English or Spanish. The authors of the ITSEA report that in
a previous study (Carter et al., 2003), the Competence Scale in par-
ticular showed high internal consistency (0.90) and strong correlations
with the Vineland Adaptive Behavior Scales for Children (Sparrow,
Ball, & Cicchetti, 1984) and the Mullen Scales of Early Learning
(Mullen, 1989).

We searched for an established measure of the developmental-
total of engagement with books and pre-literacy skills that was
appropriate for infants and toddlers, but were unable to find one.
While many measures of engagement with books (e.g., Concepts
about Print; Clay, 2006; Story and Print Concepts Tasks; Mason
& Stewart, 1989) and pre-literacy skills (e.g., Woodcock–Johnson
Letter–Word Identification; Woodcock et al., 2001; Phonological
Awareness Literacy Screening, Pre-K edition; Invernizzi, Sullivan,
Meier, & Swank, 2004) exist, they are not appropriate for use with
infants and toddlers. We developed The Apples (Moreno, 2006) for
this study to fill that void. Since the LTR-CAR is focused, in part,
on age-appropriate engagement with books, print materials, and writing and drawing materials, we compiled a checklist of solicited behaviors that would be employed by the research assistants during the home visits in order to evaluate the cross-reporter and cross-context validity of this important category of readiness skills.

Although the results for this criterion measure must be interpreted with special caution since The Apples has not yet been validated, we adapted or derived items from existing research and assessment tools. Moreover, we felt it important to include the objective perspective of the research assistants not just in a standardized format, but also in a more naturalistic context that is more consistent with the spirit of authentic assessment.

Some of the fine motor items for The Apples (which are precursors of writing skills) were adapted from the Mullen Scales of Early Learning (Mullen, 1989), such as, “grasps peg,” and “initiates a crayon stroke.” Pre-reading and book behavior items were specified, in part, according to a study by Fletcher, Perez, Hooper, and Claussen (2005) in which the investigators attempted to enhance the book reading behaviors of toddlers. The items designed for The Apples include frequency counts of behaviors such as, “turned pages in the book,” “pointed at pictures in the book,” “made eye contact with the adult during book reading,” and “looked in the same direction on the page where adult was reading/pointing (without explicit direction from adult to look there).” Children read a different book with the experimenter and the parent (the books were available in Spanish), and these episodes were scored separately. All items were summed (frequencies of passed items in the fine motor/pre-writing section and frequencies of book behavior/pre-reading items during the parent and experimenter reading sessions) to achieve one summary score. The internal consistency (Cronbach’s alpha) of this summary score was .80 in the fall and .82 in the spring.

Parental literacy. Parental literacy was assessed in this study as a potential covariate because it has been shown to be an important predictor of children’s literacy in low-income populations (e.g., Storch & Whitehurst, 2001). It provides a measure of variation in the home lives of the children in this sample, which is predominantly low-income and has low parental education. Parents’ literacy was assessed using the Letter-Word Identification subtest of the Woodcock–Johnson III Tests of Achievement (Woodcock et al., 2001) or the Batería III Woodcock–Muñoz (Muñoz-Sandoval et al., 2005). This subtest measures the respondent’s ability to identify words, including providing the correct pronunciation of words, but does not require that the individual demonstrate any knowledge of the meaning of the words. This assessment was included as a control variable.

The Learning Through Relating Child Assets Record. The LTR-CAR consists of a behavioral checklist of 144 items, as well as the Triple-A Planning Tool (Assets, Actions, and Activities), which provides the structure for documenting the authentic observations of children that provided the basis for the highest score. The tool is divided into five cross-domain topics: Play, Interacting in Tune, Expressive Communication, Receptive Comprehension, and Emerging Literacy. The research base for the LTR-CAR items, especially in the language-related topics, relied heavily on the work of Rossetti (1990) and the Hanen Centre (e.g., Weitzman & Greenberg, 2002).

Each topic has three to four skill-waves for a total of 18. A skill-wave is a developmental trajectory of eight increasingly sophisticated behavioral indicators related to the concept in the skill-wave. For example, the skill-waves under Play are: Symbolic Development, Social Game Playing, and Persistence & Exploration. As discussed earlier, the eight chronological age bands are not written on the document itself. However, the levels were intended to correspond to the following, unequal age bands: 0–1 month, 2–4 months, 5–8 months, 9–12 months, 13–18 months, 19–24 months, 25–30 months, and 31–36 months. These ranges were chosen to reflect the more rapid rate of development in the earlier part of the 0–3 age-range. The complete list of skill-waves, as well as a full example of one of them, is provided in Appendix A.

While caregivers were trained and coached on a variety of methods for “real-time” and ongoing documentation of observations throughout the three to four month observation period, the tool includes a structure for documenting the most recent examples that justify the highest Got It score in the skill-wave. Each skill-wave is to include three such examples, and therefore there are 54 anecdotes submitted with each completed LTR-CAR. Authenticity and detail in these anecdotes is an ongoing subject of reflection in coaching and booster sessions. The remainder of the Triple-A Planning Tool relates to the use of the overall LTR curriculum, where caregivers denote the plans for specific learning opportunities for each child based on the next phase in each skill-wave.

### 6. Results

#### 6.1. Descriptive statistics

Descriptive statistics for the criterion measures and LTR-CAR are displayed in Table 1, and the bivariate relationships among all major study variables are displayed in Table 2. The LTR-CAR scores showed reasonable variability, and means and ranges that corresponded approximately with the chronological ages of 0–12 month-olds, one-year-olds, and two-year-olds. Furthermore, the LTR-CAR had significant bivariate associations in the expected direction with study variables except the spring LTR-CAR scores were not associated with PLS standard scores.

#### 6.2. Reliability

As shown in Table 3, the LTR-CAR scores showed adequate reliability. Alphas were computed at both fall and spring within each age band (e.g., an alpha was computed using the 18 items that represent the 0–1 month age band). Recall that internal consistency within skill-waves (rather than age bands) would not make sense to assess because internal consistency is appropriate for a group of items that can be viewed as homogeneous (Pedhazur & Schmelkin, 1991). Since the items across the skill-waves are intended to get harder, they cannot be viewed as homogeneous. Children, particularly younger children, would not be expected to be rated similarly across the items from one skill-wave. The average alpha was .91 (range .66–.97). Reliability was lower for the younger age bands.

Table 1

<table>
<thead>
<tr>
<th>Measure</th>
<th>Mean (SD)</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>PLS Age Equivalent Fall</td>
<td>24.09 (11.47)</td>
<td>4.00</td>
<td>50.00</td>
</tr>
<tr>
<td>PLS Age Equivalent Spring</td>
<td>27.41 (10.91)</td>
<td>8.00</td>
<td>49.00</td>
</tr>
<tr>
<td>PLS Standard Score Fall</td>
<td>100.30 (16.63)</td>
<td>54</td>
<td>133</td>
</tr>
<tr>
<td>PLS Standard Score Spring</td>
<td>101.86 (16.76)</td>
<td>65</td>
<td>150</td>
</tr>
<tr>
<td>CS-ITSE Fall</td>
<td>1.42 (.35)</td>
<td>.51</td>
<td>2.00</td>
</tr>
<tr>
<td>CS-ITSE Spring</td>
<td>1.53 (.28)</td>
<td>.68</td>
<td>2.00</td>
</tr>
<tr>
<td>Apples Fall</td>
<td>38.47 (21.05)</td>
<td>3.00</td>
<td>99.33</td>
</tr>
<tr>
<td>Apples Spring</td>
<td>42.46 (19.99)</td>
<td>10.00</td>
<td>80.80</td>
</tr>
<tr>
<td>LTR-CAR Fall</td>
<td>5.08 (1.58)</td>
<td>1.92</td>
<td>7.93</td>
</tr>
<tr>
<td>0–12 Months (n = 27)</td>
<td>3.31 (1.23)</td>
<td>1.92</td>
<td>6.66</td>
</tr>
<tr>
<td>13–24 Months (n = 49)</td>
<td>5.04 (1.28)</td>
<td>2.15</td>
<td>7.93</td>
</tr>
<tr>
<td>25–38 Months (n = 59)</td>
<td>5.98 (1.19)</td>
<td>2.78</td>
<td>7.93</td>
</tr>
<tr>
<td>LTR-CAR Spring</td>
<td>6.37 (1.53)</td>
<td>2.05</td>
<td>8.00</td>
</tr>
<tr>
<td>0–12 Months (n = 13)</td>
<td>3.51 (.38)</td>
<td>2.96</td>
<td>7.95</td>
</tr>
<tr>
<td>13–24 Months (n = 38)</td>
<td>5.90 (1.06)</td>
<td>2.05</td>
<td>8.00</td>
</tr>
<tr>
<td>25–38 Months (n = 71)</td>
<td>7.16 (1.03)</td>
<td>5.82</td>
<td>8.00</td>
</tr>
</tbody>
</table>

Notes. The Apples does not have a defined upper bound since the score reflects the sum of passed items plus frequencies of naturally occurring behaviors during book-reading sessions. LTR-CAR scores are out of eight; one point for each age-band passed.
likely because the sample had a small number of very young infants.

6.3. Internal validity

1. Do the items go in the order proposed, that is, become increasingly difficult to pass? As shown in Table 4, when percentages are pooled across all 18 items within an age-level, the items do become increasingly difficult to pass. There is also more variability (as indicated by range) as the items get harder. When items and skill-waves were examined individually (not shown), out of 144 items, there were five cases when two adjacent items were out of order. Four of five of those cases were the first two age bands, that is, between the 0–1 month item and the 2–4 month item, the pass rates were near identical (e.g., 99.4 and 99.7), and the sample sizes for children who did not pass these items was extremely low (range 1–7 children). The fifth case occurred in the Book Behavior skill-wave. The items were, “Knows how to turn pages and when to do so, when a page ends,” (intended to be 13–18 month item, actual pass rate was 64.7%) and “Spends time looking at books on own” (intended to be 19–24 month item, actual pass rate was 74.6%).

Table 4
Bivariate correlations among major study variables.

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age Fall</td>
<td>1.0</td>
<td>.94</td>
<td>.68</td>
<td>.64</td>
<td>.83</td>
<td>.82</td>
<td>−.11</td>
<td>−.12</td>
<td>.63</td>
<td>.47</td>
<td>.75</td>
<td>.64</td>
</tr>
<tr>
<td>LTR CAR Fall</td>
<td>1.0</td>
<td>.65</td>
<td>.70</td>
<td>.86</td>
<td>.17</td>
<td>−.05</td>
<td>.65</td>
<td>.51</td>
<td>.73</td>
<td>.66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LTR CAR Spring</td>
<td>1.0</td>
<td>.84</td>
<td>.79</td>
<td>.77</td>
<td>.30</td>
<td>.48</td>
<td>.63</td>
<td>.41</td>
<td>.48</td>
<td>.53</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLS Age Fall</td>
<td>1.0</td>
<td>.83</td>
<td>.80</td>
<td>.18</td>
<td>.20</td>
<td>.59</td>
<td>.35</td>
<td>.57</td>
<td>.69</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLS Age Spring</td>
<td>1.0</td>
<td>.93</td>
<td>.45</td>
<td>.32</td>
<td>.61</td>
<td>.38</td>
<td>.76</td>
<td>.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLS Standard Score Fall</td>
<td>1.0</td>
<td>.29</td>
<td>.30</td>
<td>.07</td>
<td>.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PLS Standard Score Spring</td>
<td>1.0</td>
<td>.64</td>
<td>.48</td>
<td>.59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS ITSEA Fall</td>
<td>1.0</td>
<td>.12</td>
<td>.63</td>
<td>.51</td>
<td>.73</td>
<td>.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CS ITSEA Spring</td>
<td>1.0</td>
<td>.95</td>
<td>.95</td>
<td>.95</td>
<td>.94</td>
<td>.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Apples Fall</td>
<td>1.0</td>
<td>.94</td>
<td>.96</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. The overall percentages are pooled across 18 skill-waves, and across fall and spring.

2. Do children individually increase across the academic year at a rate consistent with the amount of time that passed between assessments? This question was addressed by conducting paired t-tests of LTR-CAR scores from fall to spring. LTR-CAR scores are out of eight; one point for each age band passed. These results, for the whole sample and by broad age-groupings (infants, one-year-olds, and two-year-olds) are displayed in Table 5. As shown, children in all age groups scored significantly higher in the spring than they did in the fall. Moreover, the magnitudes of the increases were similar to the amount of time that passed between the fall and spring assessments, which was approximately nine months. The average increase of 1.81 age-levels, multiplied by the average number of months in each age-level (4.5 months) results in 8.15, or an increase of approximately one month less than the amount of time that passed.

3. Do the actual ages of children match with caregivers’ age-performance scoring? LTR-CAR scores were significantly linearly related to chronological age: Fall r(135) = .68, p < .001; Spring r(122) = .70, p < .001. In addition, this question was addressed by examining the CAR-age deviation scores (see Table 6) in a descriptive fashion. The CAR-age deviation score is the child’s actual age-group, out of nine (because some children were older than 36 months) minus the age band in which the child scored. Thus, a score of zero is a match between chronological and LTR-CAR age bands; a negative number is the number of age bands below chronological age, and a positive number is the number of age bands above chronological age. The CAR-age deviation scores reveal that, on average, children were scoring close to one full age-level below their chronological age in the fall, and virtually at age-level by spring. Given the EHS population, including a minimum of 10% of children with diagnosed delays or disabilities, a mean score of slightly below chronological age is to be expected. The CAR-age deviation scores by major age groups reveals that the greatest discrepancy between CAR scores and chronological age occurred for 2-year-olds in the fall. At this time point, this group scored more than one and a half age bands below their chronological age, on average. This discrepancy was reduced by 69% in the spring (1.61–.50/1.61). In sum, there was a trend for scores to be slightly below chronological age. This was true to a greater extent in the fall and for older children.

4. Do children get significantly closer to age-expected performance by spring? To address this question, we conducted paired t-tests on the fall-spring changes in CAR-age deviation scores from Table 6, discussed descriptively above. The CAR-age deviation score changes from fall to spring were in the desired direction and statistically significant for the whole sample, and for each age group. While all children improved, infants exhibited a unique pattern: they were the only group to perform above their actual age in the fall, and even more above their actual age in the

Table 3
Internal consistency (Cronbach’s alpha) within age-level for the LTR-CAR.

<table>
<thead>
<tr>
<th></th>
<th>Fall, n = 122</th>
<th>Spring, n = 115</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–1 Month</td>
<td>.66</td>
<td>.90</td>
</tr>
<tr>
<td>2–4 Months</td>
<td>.76</td>
<td>.89</td>
</tr>
<tr>
<td>5–8 Months</td>
<td>.93</td>
<td>.93</td>
</tr>
<tr>
<td>9–12 Months</td>
<td>.95</td>
<td>.97</td>
</tr>
<tr>
<td>13–18 Months</td>
<td>.95</td>
<td>.95</td>
</tr>
<tr>
<td>19–24 Months</td>
<td>.94</td>
<td>.97</td>
</tr>
<tr>
<td>25–30 Months</td>
<td>.94</td>
<td>.96</td>
</tr>
</tbody>
</table>

Note. Recall that each age band has 18 items.
Regressions predicting PLS-4 from LTR-CAR scores.

Table 7

<table>
<thead>
<tr>
<th>Fall, mean (SD)</th>
<th>Spring, mean (SD)</th>
<th>Difference</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole sample, n = 98</td>
<td>4.92 (1.53)</td>
<td>6.73 (1.19)</td>
<td>1.81</td>
</tr>
<tr>
<td>Infants, n = 24</td>
<td>3.43 (1.25)</td>
<td>5.66 (1.24)</td>
<td>2.23</td>
</tr>
<tr>
<td>One-year-olds, n = 39</td>
<td>5.00 (1.18)</td>
<td>6.67 (.98)</td>
<td>1.67</td>
</tr>
<tr>
<td>Two-year-olds, n = 35</td>
<td>5.86 (1.25)</td>
<td>7.53 (.68)</td>
<td>1.67</td>
</tr>
</tbody>
</table>

** * p < .01.

Table 6

<table>
<thead>
<tr>
<th>Fall (SD)</th>
<th>Spring (SD)</th>
<th>Absolute value of improvement</th>
<th>t</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole sample</td>
<td>−.86 (1.43)</td>
<td>−.05 (1.05)</td>
<td>.81</td>
</tr>
<tr>
<td>Infants</td>
<td>.28 (1.07)</td>
<td>.89 (1.05)</td>
<td>.61</td>
</tr>
<tr>
<td>One-year-olds</td>
<td>−.62 (1.35)</td>
<td>−.08 (1.07)</td>
<td>.54</td>
</tr>
<tr>
<td>Two-year-olds</td>
<td>−1.61 (1.23)</td>
<td>−.50 (.69)</td>
<td>1.11</td>
</tr>
</tbody>
</table>

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

6.4. External validity

External validity was analyzed in two ways: (a) a series of regressions using LTR-CAR scores, after accounting for major covariates, to predict three different criterion measures (with the PLS scores reported in two ways); and (b) cross-tab correspondence between LTR-CAR and the criterion measures on at-risk vs. not-at-risk, based on standard deviation cut-offs. Prior to conducting regressions, analyses were conducted to assess the extent to which the covariates were associated with the LTR-CAR. A series of t-tests was conducted for both ethnicity (minority vs. non-minority) and gender, with LTR-CAR scores and all criterion measures as dependent variables. None of the ethnicity t-tests was significant, and therefore this variable was dropped as a covariate from further analyses. Gender was not associated with PLS-4 or Apples scores, but girls had significantly higher LTR-CAR scores in the fall; M = 6.78 (SD = 1.12) for girls, M = 5.89 (SD = 1.75) for boys, t(101) = 3.04, p < .01, and girls also had significantly higher CS-ITSEA scores in the spring; M = 1.46 (SD = .30) for girls, M = 1.29 (SD = .37) for boys, t(58) = 2.11, p < .05. This is probably due to the fact that girls were slightly older. This difference missed statistical significance in the fall (20 vs. 23 months, p = .06) but was significant in the spring (25 vs. 29 months, p < .05). Thus, gender and age were retained as covariates. Interestingly, parental literacy (Woodcock-Johnson Letter Word Recognition) was not correlated with LTR-CAR scores or any of the criterion measures with the exception of PLS-4 standard score in the spring, t(57) = .31, p < .05, and therefore this variable was not retained as a covariate.

The results for the regressions predicting PLS scores are reported in Table 7 and the results for the regressions predicting CS-ITSEA and Apples scores are reported in Table 8. For all eight regressions, the model containing LTR-CAR score as a predictor (Step 2) was statistically significant, accounting for between 14% and 75% of the variance in the criterion measure (average R² = .50). However, in two cases, CS-ITSEA in the spring and Apples in the fall, chronological age subsumed the majority of variance, and the LTR-CAR did not contribute a significant amount of unique variance. For the six regressions in which LTR-CAR’s unique variance was statistically significant over and above age and gender, it ranged from 6% to 26% (average 11%). In order to establish and cross-tabulate “at-risk” and “not at-risk” groups, cut-offs were established for LTR-CAR score and all
criterion measures. For all measures except PLS standard score, the cut-off was one standard deviation below the mean. Children who scored one standard deviation below the mean or higher were placed in the not at-risk category and children who scored less than one standard deviation below the mean were placed in the at-risk category. For PLS standard score, the cut-off was 1.5 standard deviations below the mean, in accordance with standardization procedures reported in the technical manual for this assessment.

As shown in Table 9, there was high overall agreement on true negatives and true positives combined (range, 77.6–89.1%, average agreement, 84.6%). Once again, the performance of the PLS standard score stood out as different from the other criterion indices. Despite an overall high level of agreement even with this index, the correlation between it and the LTR-CAR was not statistically significant, likely due to the overall low number of children who fell into the at-risk category for the PLS standard score (3 and 4, for fall and spring respectively). In other words, the LTR-CAR seems to be a more stringent assessment of delay than the PLS-4 standard score (i.e., 19.4% of children in the fall and 9.1% in the spring for which the LTR-CAR categorized the child at risk but the PLS standard score did not) because the rate of false negatives was lower. This is consistent with the PLS’ own validation study, using matched samples of children with diagnosed language disorders. In a small sample of 48 children who were ages 3 to 3 years 11 months (the youngest group studied), PLS had a 10.4% false negative rate and 4.2% false positives.

False negatives are more concerning than false positives because they present a risk that a child who needs additional support may not get it. LTR-CAR’s mean false positive rate was 7.6% and mean false negative rate was 7.8%. The false negative rate was driven up primarily by the CS-ITSEA and Apples, both in the spring. Thus, the LTR-CAR was somewhat more conservative than the PLS and somewhat less conservative than the ITSEA and Apples. All in all, the lower false negative rate for the LTR-CAR than the PLS had with its own criterion reference when it was validated is reassuring given that the most likely concern would have been that infant–toddler caregivers may have the tendency to be too lenient in their scoring.

6.5. Exploratory analyses of center-based vs. home-based caregivers

We could not analyze for differences in the strength of the relationship between LTR-CAR and the criterion measures by caregiver level of education and child care experience since the tools are filled out by teaching teams, with the exception of home visitors. However, since the three home visitors did fill out the CAR independently (not including coach assistance), we could analyze for differences by center-based caregivers vs. home-based caregivers, as well as different average levels of education across these groups, if any. This was important to do, not only because of potential background differences in those who might be attracted to these different job types, but also because no similar evidence exists to show that a normative infant–toddler authentic assessment is employable in an exclusively home-based service model. Given that home visitors spend dramatically less time with children than do center-based teachers, it is a real concern whether home visitors can provide equally accurate descriptions of children’s development.

Table 8

<table>
<thead>
<tr>
<th></th>
<th>CS-ITSEA</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fall</td>
<td>Spring</td>
<td>Fall</td>
</tr>
<tr>
<td>B</td>
<td>β</td>
<td>B</td>
<td>β</td>
</tr>
<tr>
<td>Step 1</td>
<td>Gender</td>
<td>.13</td>
<td>.18</td>
</tr>
<tr>
<td>Age</td>
<td>.02</td>
<td>.58</td>
<td>.02</td>
</tr>
<tr>
<td>Model</td>
<td></td>
<td>F</td>
<td>R²</td>
</tr>
<tr>
<td></td>
<td>22.02</td>
<td>.40**</td>
<td>15.17</td>
</tr>
<tr>
<td>Step 2</td>
<td>Gender</td>
<td>.13</td>
<td>.18</td>
</tr>
<tr>
<td>Age</td>
<td>.01</td>
<td>.33**</td>
<td>.02</td>
</tr>
<tr>
<td>LTR-CAR</td>
<td>.09</td>
<td>.37**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>F</td>
<td>R²</td>
<td>F</td>
</tr>
<tr>
<td></td>
<td>19.31</td>
<td>.48**</td>
<td>10.07</td>
</tr>
<tr>
<td>LTR-CAR unique variance</td>
<td>8.68</td>
<td>.07**</td>
<td>4.8</td>
</tr>
</tbody>
</table>

Notes. The reference category is at-risk status. Thus a positive classification occurs when a child falls more than one standard deviation below the mean.

6.6. Exploratory analyses of center-based vs. home-based caregivers

We could not analyze for differences in the strength of the relationship between LTR-CAR and the criterion measures by caregiver level of education and child care experience since the tools are filled out by teaching teams, with the exception of home visitors. However, since the three home visitors did fill out the CAR independently (not including coach assistance), we could analyze for differences by center-based caregivers vs. home-based caregivers, as well as different average levels of education across these groups, if any. This was important to do, not only because of potential background differences in those who might be attracted to these different job types, but also because no similar evidence exists to show that a normative infant–toddler authentic assessment is employable in an exclusively home-based service model. Given that home visitors spend dramatically less time with children than do center-based teachers, it is a real concern whether home visitors can provide equally accurate descriptions of children’s development.

Table 9

<table>
<thead>
<tr>
<th></th>
<th>PLS age equivalent</th>
<th></th>
<th></th>
<th></th>
<th>CS-ITSEA</th>
<th></th>
<th></th>
<th>Apples</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Fall</td>
<td>Spring</td>
<td>Fall</td>
<td>Spring</td>
<td>Fall</td>
<td>Spring</td>
<td>Fall</td>
<td>Spring</td>
<td>Fall</td>
<td>Spring</td>
</tr>
<tr>
<td>n=65</td>
<td>86.1</td>
<td>89.1</td>
<td>77.6</td>
<td>83.6</td>
<td>88.3</td>
<td>82.4</td>
<td>84.4</td>
<td>85.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>91.1</td>
<td>89.1</td>
<td>77.6</td>
<td>83.6</td>
<td>88.3</td>
<td>82.4</td>
<td>84.4</td>
<td>85.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=55</td>
<td>12.3</td>
<td>9.3</td>
<td>1.5</td>
<td>0.0</td>
<td>11.8</td>
<td>3.5</td>
<td>9.4</td>
<td>7.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>80.0</td>
<td>80.0</td>
<td>76.1</td>
<td>83.6</td>
<td>76.5</td>
<td>78.9</td>
<td>75.0</td>
<td>77.8</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=67</td>
<td>73.8</td>
<td>9.0</td>
<td>19.4</td>
<td>9.1</td>
<td>8.8</td>
<td>3.5</td>
<td>10.9</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=55</td>
<td>9.2</td>
<td>0.0</td>
<td>19.4</td>
<td>9.1</td>
<td>8.8</td>
<td>3.5</td>
<td>10.9</td>
<td>0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=68</td>
<td>4.6</td>
<td>10.9</td>
<td>3.0</td>
<td>7.3</td>
<td>2.9</td>
<td>14.0</td>
<td>4.7</td>
<td>14.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>85.9</td>
<td>85.9</td>
<td>85.9</td>
<td>85.9</td>
<td>85.9</td>
<td>85.9</td>
<td>85.9</td>
<td>85.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=57</td>
<td>.56**</td>
<td>.63**</td>
<td>.07</td>
<td>-.09</td>
<td>.61**</td>
<td>.23**</td>
<td>.47**</td>
<td>.53**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=54</td>
<td>.56**</td>
<td>.63**</td>
<td>.07</td>
<td>-.09</td>
<td>.61**</td>
<td>.23**</td>
<td>.47**</td>
<td>.53**</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes. The reference category is at-risk status. Thus a positive classification occurs when a child falls more than one standard deviation below the mean.

*p < .05.

**p < .01.

†p < .10.
In analyzing years of formal education, length of time in the field overall, and length of time working at this particular center, we found that all of the means favored the home visitors over center-based providers (that is, home visitors had more education and experience) but none significantly so. Thus, any differences seen in the strength of the criterion validity relationships between home visitors and center-based teachers could not be due to home visitors being less educated or experienced than center-based or combination teachers, but instead might be due to the decreased observation time available during once weekly home visits.

We re-ran the eight regressions separately for center-based (n = 38) and exclusively home-based children (n = 29) with complete data, and examined only the unique variance contributed by the LTR-CAR at Step 2. Although three of the regressions changed from significant to non-significant once the sample was split, the differences did not favor either group: In two cases, both unique variances were still statistically significant (PLS age equivalent fall and PLS standard score fall); in three cases the center-based teachers’ LTR-CAR scores contributed slightly more unique variance (though non-significant) and in the remaining three cases, the home visitors’ LTR-CAR scores contributed slightly more unique variance (though non-significant). In sum, the results support an initial conclusion that home visitors and center-based caregivers do not differ in the strength of predictions from the LTR-CAR to the criterion measures.

7. Discussion

This study examined the reliability and validity of scores obtained by non-specialized infant–toddler caregivers on a new authentic assessment, the Learning Through Relating Child Assets Record. We examined internal consistency, expected “internal validity” performance (e.g., that scores should be related to chronological age, items should go in the order of difficulty proposed), and “external validity” performance in terms of its concurrent associations with three criterion measures. Furthermore, LTR-CAR data as well as data on the criterion measures were collected at both fall and spring, and therefore developmental change captured by the measures was also analyzed.

In terms of descriptive statistics, reliability, and internal validity, the evidence seems to support a strong internal structure of the LTR-CAR that was employable by these infant–toddler caregivers. It is possible that internal consistency was somewhat inflated by the “all children–all items” format of the instrument. Scores were strongly related to chronological age despite the fact that the age bands are not mentioned on the measure or any materials that the caregivers receive. Although the stimulus properties around the order of difficulty of items were obvious and discussed openly in training and coaching, it was also a part of training and coaching that caregivers were allowed and encouraged to score items out of order if that was how a child was performing. Anecdotally, we know that at least some teachers on some occasions were willing to do this; the aggregate results bore this out in only one notable case. It is not surprising, perhaps, that this happened in the “Book Behavior” skill wave – the least well enumerated developmental trajectory in previous research.

Also notable in terms of internal structure and functionality of the LTR-CAR is that scores followed the expected pattern for this Early Head Start population, which bears higher than average levels of risk. There was a trend for children to score below their chronological age in the fall, and this negative discrepancy was larger for older children. Within-subjects analyses of fall-spring changes in CAR-age deviation scores indicated significant improvement for the whole sample and for each of infants, one-year-olds, and two-year-olds. Although participation in an Early Head Start program is intended to help improve developmental outcomes for at-risk children, the effect sizes seen here are larger than have been found in experimental research (e.g., Love et al., 2005). We note that regression to the mean could partially explain these seemingly large effects of “catching up.” Moreover, EHS found child impacts on standardized, not criterion-referenced, assessments such as the Bayley Scales (Bayley, 1993) and the Peabody Picture Vocabulary Test (Dunn & Dunn, 2007), and the comparisons were between subjects of the same age (e.g., 36 months). A larger “impact,” therefore, is expected in the current case because the LTR-CAR has specific behavioral referents that teachers are explicitly trained to support, which we analyzed for within children. We find it unlikely that this effect could be explained by socially desirable responding by the teachers in the spring, but not in the fall. This is especially implausible given the fact that children initially performing below chronological age in the fall remained a non-trivial amount below chronological age (one third of an age band) in the spring.

Actual accuracy of caregivers’ LTR-CAR ratings was assessed relative to three criterion measures: The PLS-4, standard and age-equivalent scores, the Competence Scale of the ITSEA, and The Apples, an observer assessment of naturalistic pre-reading and pre-writing skills designed for this study. Although The Apples is not a validated measure, we include it as an additional reference point for consideration because of the centrality of the skills it measures to the LTR system, and because it allowed for the viewpoint of objective research assistants during a semi-naturalistic playing and reading session. In all, we compared the viewpoints of the children’s professional caregivers, parents, and dispassionate observers in both standardized and semi-naturalistic formats.

Although the regression models predicting the criterion measures from LTR-CAR scores indicated that chronological age subsumed a large amount of these associations, in only two out of eight cases (CS-ITSEA in the spring and Apples in the fall) did the LTR-CAR not account for a significant amount of additional unique variance. When you consider both the differences in reporter and context between the LTR-CAR and the criterion measures, and the fact that age and authentic assessments are expected to be strongly (but not perfectly) linearly related, the resulting contribution of the LTR-CAR was quite striking. Particularly in the case of the PLS, arguably the most credible criterion measure used here and most different from the LTR-CAR, the associations were consistent across both standard and age-equivalent scores, and the percent of variance accounted for by the LTR-CAR ranged from 6% to 26% (average 11%).

These relationships were further explored by dichotomizing the LTR-CAR and criterion measures, and examining agreement regarding children’s at-risk status. On average, agreement on classification status was high overall (84.6%). Particularly in the fall, the LTR-CAR scored more children as at-risk than the PLS standard score did. Given that the greatest concern is likely that infant–toddler caregivers may be reluctant to indicate that a child cannot do something, or have difficulty seeing delay or challenge in children with whom they have close relationships (Meisels et al., 2010), the generally low rate of false negatives (not identifying a child as at-risk when she should have been), 7.8% on average, reduces this concern. This rate of false negatives is lower than the PLS-4’s own rate, which hovered around 10%, for its validation sub-study (Zimmerman et al., 2002).

Finally, given the dramatic differences in the amount of time home visitors spend with children compared to center-based teachers, it was important to assess whether home visitors were less accurate in their ratings. This question was analyzed in an exploratory fashion, given the reduction in variance and sample size when splitting the sample this way. Some of the results changed from significant to non-significant, and it is not prudent to make strong conclusions from non-significant results. How-
ever, given that the two remaining significant regressions were significant for both home- and center-based caregivers, and the remaining non-significant regressions favored each group equally, we can tentatively suggest that there is no strong evidence to believe that the home visitors in this sample had additional difficulty in achieving accuracy over the center-based caregivers.

Further study, including qualitative assessments, is needed to assess the unique challenges and barriers for home visitors in using the LTR-CAR, and to establish whether the use of the LTR-CAR could be employable by home visitors in other contexts.

Two limitations stand out as the most significant threats to the certainty of conclusions made here. First and foremost, this study took place in only one Early Head Start program. Early Head Start programs do not necessarily map on to other types of infant–toddler programs, and in some known ways, do not. For example, Head Start and Early Head Start standards, which are established federally, have higher education requirements for teachers than most state licensing requirements. The fact that this was an Early Head Start program alone likely explains at least part of the seemingly high ability of these caregivers to employ a 144-item authentic assessment. More importantly, every EHS center is structurally and functionally unique, and thus the ability to generalize the usability of the LTR-CAR to other locations is extremely limited. Although national statistics on the education backgrounds of the infant–toddler child care work force are not available, it is estimated that approximately 80% have some college or more (Day, 2006). The same statistic for the current sample is 83%. However, the National Association of Child Care Resources & Referral Agencies (NACCRRA, 2010) estimates that 33% of all center-based child care teachers have graduated from college (http://www.naccra.org). Given that this statistic is 37% for the current sample, and that the current sample is exclusively infant–toddler, it is probably safe to conclude that the formal education level of the caregivers in this sample is likely somewhat higher than the overall infant–toddler caregiver workforce.

The second significant limitation on the certainty of results presented here is the very small number of very young infants enrolled in this program. Although we explored the internal structure of the LTR-CAR by major age sub-group, i.e., infants, one-year-olds, and two-year-olds, the sample sizes for the individual infant age bands were prohibitively small to conduct any analyses that could discriminate among the items at this micro level (e.g., two 0–1 month-olds in the fall and none in the spring; eight 2–4 month-olds in the fall and two in the spring). This was another consequence of having conducted the study at only one center: we made no attempt to influence recruitment or enrollment at the program. Thus, now that the LTR-CAR shows promise for usability and usefulness, it must be implemented and researched in additional locations in order to assure that the items are usable and accurate for the very youngest of the birth-to-three range.

Our decision to not implement LTR beyond the original development site, despite many programs requesting to use it, was quite intentional. Because of the investment required by the program (i.e., training and ongoing support, approximately one year of technical assistance to implement the LTR-CAR) and significant change in practice that is expected in order to use the LTR system with fidelity, we felt it appropriate to place the initial onus of validation on the development site, rather than place demands on multiple sites in exchange for benefits we could not yet promise. In other words, we decided to take a very conservative approach to dissemination in order not to add to the confusion that exists around the evidence-base for child assessments, especially for the youngest children. Yet, the data contained herein, including standardized and valid criterion measures at two time-points, make an important contribution to the literature given the dearth of evidence-based assessment options particularly at the infant–toddler level. The unique features of the LTR-CAR (e.g., no mention of chronological age, the “all children–all items” format, the non-traditional structure of skill-waves into generalized learning “backdrops” rather than domains of development) were chosen to better reflect the way that infants and toddlers learn and act upon their world. Nevertheless, it was a real concern whether some of these features might have led to difficulty in implementation. Thus, the internal structure, as well as associations with external measures evidenced here, provide initially encouraging evidence that the LTR-CAR could provide a viable option for a developmentally appropriate authentic assessment, employable by infant–toddler teachers with limited training.

As we begin to work with additional programs to implement LTR, it is vitally important that our next research priority is to examine the impact of the use of the LTR-CAR on, first, teacher practice, and second, child outcomes. The stated purpose of authentic assessment is not to categorize children, or grant or deny services, but to plan for appropriate learning opportunities. Although we collected observational, interview, and self-report data on caregivers’ use of LTR overall, we have not yet implemented a design that would allow a direct link between quality of LTR-CAR implementation and desired changes in the learning environment. For example, linking qualitative analysis of the authenticity and depth of caregivers’ documentation anecdotes to a validated observational assessment of teacher–child interactions would provide evidence for whether using the LTR-CAR effectively leads to day-to-day improved individualization of instruction. As researchers in partnership with community early childhood programs, we are mutually interested in authentic assessment only insofar as it serves as a tool in the arsenal for improving the quality of early learning experiences.
### Appendix A. Skill-wave titles and example of a full skill-wave

<table>
<thead>
<tr>
<th>Topic</th>
<th>Skill-waves</th>
</tr>
</thead>
</table>
| Play                          | 1. Symbolic Development
                                  | 2. Social Game Playing
                                  | 3. Persistence and Exploration
                                  | 4. Joint Attention
                                  | 5. Affect Sharing
                                  | 6. Self-regulation
                                  | 7. Peer Play and Social Relationships |
| Interacting in Tune           |                                                                           |
| Expressive Communication      | 8. Sounds and Words in Context
                                  | 9. Rules of Conversation
                                  | 10. Expresses Protest
                                  | 12. Responsive Interacting
                                  | 13. Understands Others' Feelings
                                  | 14. Understands Routines
                                  | 15. Recognizes and Identifies Objects |
| Receptive Communication       |                                                                           |
| Emerging Literacy             | 16. Book Behavior
                                  | 17. Pre-reading
                                  | 18. Pre-Writing |

<table>
<thead>
<tr>
<th>Items in Social Game Playing Skill-wave (age bands included for clarity; these are not printed in the assessment)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0–1 Month: Maintains eye contact for several seconds</td>
</tr>
<tr>
<td>2–4 Months: Smiles at person who tries to get his attention</td>
</tr>
<tr>
<td>5–8 Months: Takes turns in games – urges the adult to continue the game through gestures, facial expressions or vocalizations</td>
</tr>
<tr>
<td>9–12 Months: Attempts to initiate combinations of actions in games (e.g., a facial expression, a noise, and a gesture)</td>
</tr>
<tr>
<td>13–18 Months: Initiates a turn-taking routine with an adult, i.e., starts a game</td>
</tr>
<tr>
<td>19–24 Months: Participates appropriately in a social game (over several turns) that involves several modalities (face, voice, gesture) and/or objects</td>
</tr>
<tr>
<td>25–30 Months: Elaborates upon a social game that involves several modalities and/or objects (intentionally adds new actions)</td>
</tr>
<tr>
<td>31–36 Months: Initiates and directs a social game that involves several modalities and/or objects</td>
</tr>
</tbody>
</table>

### References