

Published in final edited form as:

*Early Child Res Q.* 2013 April 1; 28(2): 218–233. doi:10.1016/j.ecresq.2012.12.004.

## New evidence on the validity of the Arnett Caregiver Interaction Scale: Results from the Early Childhood Longitudinal Study-Birth Cohort

Nicole Colwell<sup>a,\*</sup>, Rachel A. Gordon<sup>b,c</sup>, Ken Fujimoto<sup>a</sup>, Robert Kaestner<sup>c,d</sup>, and Sanders Korenman<sup>e</sup>

<sup>a</sup>Department of Educational Psychology, University of Illinois at Chicago, United States

<sup>b</sup>Department of Sociology, University of Illinois at Chicago, United States

<sup>c</sup>Institute of Government and Public Affairs, University of Illinois at Chicago, United States

<sup>d</sup>Department of Economics, University of Illinois at Chicago, United States

<sup>e</sup>School of Public Affairs, Baruch College/CUNY, New York, NY, United States

### Abstract

The Arnett Caregiver Interaction Scale (CIS) has been widely used in research studies to measure the quality of caregiver–child interactions. The scale was modeled on a well-established theory of parenting, but there are few psychometric studies of its validity. We applied factor analyses and item response theory methods to assess the psychometric properties of the Arnett CIS in a national sample of toddlers in home-based care and preschoolers in center-based care from the Early Childhood Longitudinal Study-Birth Cohort. We found that a bifactor structure (one common factor and a second set of specific factors) best fits the data. In the Arnett CIS, the bifactor model distinguishes a common substantive dimension from two methodological dimensions (for positively and negatively oriented items). Despite the good fit of this model, the items are skewed (most teachers/caregivers display positive interactions with children) and, as a result, the Arnett CIS is not well suited to distinguish between caregivers who are “highly” versus “moderately” positive in their interactions with children, according to the items on the scale. Regression-adjusted associations between the Arnett CIS and child outcomes are small, especially for preschoolers in centers. We encourage future scale development work on measures of child care quality by early childhood scholars.

### Keywords

Child care quality; Arnett Caregiver Interaction Scale; Factor analysis; Item response theory; Validity

### 1. Introduction

As use of non-parental child care increased during the latter half of the twentieth century, scales of caregiver-child interactions emerged. To a varying degree, these scales drew on developmental theory, which suggests that caregiver-child interactions are important influences on children’s cognitive and socio-emotional development. Jeffrey Arnett

---

\* Corresponding author at: Institute of Government and Public Affairs, University of Illinois at Chicago, 815 West Van Buren St., Suite 525, Chicago, IL 60607, United States. Tel.: +1 312 996 6188; fax: +1 312 996 1404. ncolwe2@uic.edu (N. Colwell).

developed one widely used scale, the Arnett Caregiver Interaction Scale (CIS; Arnett, 1986), for an evaluation of a caregiver training program at Bermuda College. Arnett described writing items to encompass the parenting constructs of demandingness, responsiveness, and explanations for punishment based on four well-established parenting styles: authoritarian, authoritative, permissive, and uninvolved (Arnett, 1986, 1989; Baumrind, 1967; Maccoby & Martin, 1983). In a study of 66 caregivers in 22 centers in Bermuda, Arnett used an exploratory principal component analysis and selected a four-dimensional structure (see Table 1).

The Arnett CIS has since been used in numerous studies of child care including recent, large-scale studies such as the Early Childhood Longitudinal Study-Birth Cohort (ECLS-B; Mollborn & Blalock, 2012); Early Head Start Research and Evaluation Project (Love et al, 2004); Head Start Family and Child Experiences Surveys (FACES; Bracken & Fischel, 2006); and, Welfare, Children and Families: A Three-City Study (Votruba-Drzal, Coley, Maldonado-Carreo, Li-Grining, & Chase-Lansdale, 2010). Despite its widespread use, there have been few studies of the psychometric properties of the Arnett CIS. As we review below, only a handful of factor analyses have been reported and we are aware of no prior item response theory analyses of the scale. Our study contributes a more comprehensive study of the validity of the scale with a large sample of caregivers from across the nation. The study is intended to guide researchers conducting secondary analyses of these numerous large, public datasets as well as scholars contemplating using the scale in future data collection efforts regarding child development, child care, and early childhood education.

### 1.1. Review of Prior Factor Analyses and Predictive Validity of the Arnett CIS

It is surprising that the Arnett CIS has been so widely adopted in child care research in the U.S. given that it was developed by a doctoral student for a small-scale evaluation outside of the U.S. Its popularity likely reflects its grounding in one of the most influential parenting frameworks in the developmental literature, its face validity connection to that framework, and the fact that Arnett's dissertation was part of a broader research initiative in Bermuda led by preeminent child care scholars (McCartney, Scarr, Phillips, & Grajek, 1985; Scarr & McCartney, 1988). In fact, other than the addition of examples, the wording of the scale has remained largely intact since Arnett first published the scale and, as noted above, it has been incorporated into many large-scale studies of child care and preschool.

Reflecting the standard of practice at the time, Arnett did not use contemporary measurement approaches to define dimensions and write items. Instead, he empirically identified dimensions using a principal component analysis. Table 1 presents the resulting four dimensions (factors) and items for each dimension and shows the slight modifications made in later studies. Namely, two factors have been relabeled: the Sensitivity dimension was called "Positive Relationship" by Arnett; the Harshness dimension was called "Punitiveness." One item is positioned differently: Arnett placed Item 24 ("Expects the child to exercise self-control") on the fourth factor (Permissiveness) whereas Table 1 follows later studies and places it with Harshness. Additionally, as can be seen in comparing the third and fourth columns of Table 1, later studies expanded item wording, generally adding examples of behaviors consistent with a higher score (Snow et al, 2007).

Few replications of Arnett's factor structure have been published in scholarly journals. Those that exist are found in technical reports from large-scale studies. Results from these studies showed that Arnett's first three factors are better replicated than the fourth factor (Permissiveness), which may reflect the fact that just three (often quite skewed) items comprised the final factor. The best documented factor analyses are found in reports from Abt Associates (Layzer, Goodson, & Moss, 1993) and RMC Research Corporation (Love, Ryer, & Faddis, 1992) to state and federal education departments. Abt Associates' report

was based on a study of 119 centers, preschools and Head Start programs that served low-income children (Layzer et al., 1993). Their factor analysis of the Arnett CIS produced a five-factor solution (Layzer et al., 1993). The first three factors largely replicated Arnett's structure (the same items loaded on the Sensitivity and Detachment factors shown in Table 1; all but one of the same items loaded on the Harshness factor). The fourth and fifth factors differ from Arnett's original structure (items 4 and 18 from Table 1 loaded on Layzer and colleagues' fourth factor; items 9 and 15 loaded on a fifth factor). The California Staff/Child Ratio study examined associations of child/staff ratios with quality of care in a sample of 122 classrooms across the San Francisco, Los Angeles and San Diego areas (Love et al., 1992). It reported four factors in the Arnett CIS items that accounted for 60% of the variance. Like the authors of the Abt report, Love and colleagues' first three factors were similar to Arnett's original structure with the exception of one of the Harshness items. In contrast, their fourth factor included items 4, 9, 18 and 24. Love and colleagues also deleted item 15, which failed to load on any factor.

Other reports provide only the number of factors, but not item loadings. In their report associated with the National Child Care Staffing Study, Whitebook, Howes, and Phillips (1989) found three factors accounting for 60% of the variance which they labeled "sensitivity," "harshness," and "detachment." The Early Head Start Research and Evaluation Project and FACES studies relied on a single summary score based on an average of all 26 Arnett CIS items, because factor analyses did not reveal distinct dimensions (Love et al., 2004) and because the subscales were highly inter-correlated (Zill et al., 2006). In analyses of a Dutch sample, van IJzendoorn, Tavecchio, Stams, Verhoeven, and Reiling (1998) identified a twofactor solution in which half of the 26 items loaded on each factor; they termed the factors "authoritarian interaction" and "stimulating interaction."

Academic publications that use the Arnett CIS generally rely on Arnett's original structure, or one of the modifications reviewed above, sometimes reporting internal consistency reliability coefficients, but not factor analytic confirmations of the structure. For example, Bracken and Fischel (2006) & Maxwell, McWilliam, Hemmeter, Ault, and Schuster (2001) referred to Arnett's original four-factor structure in their studies. Bracken and Fischel reported a Cronbach's alpha of .94 for all items; Maxwell et al. (2001) reported an average Cronbach's alpha of .93. In contrast, Cryer, Tietz, Burchinal, Leal, and Palacios (1999) & Tietz, Cryer, Bairrao, Palacios, and Wetzel (1996) relied on Whitebook et al.'s (1989) three-factor structure and reported Chronbach's alpha ranges of .85-.94 for sensitivity, .39-.78 for involvement/detachment, and .83-.93 for acceptance/harshness. Torquati, Raikes, and Huddleston-Casas (2007) did not identify the source of their factors, citing Arnett's 1989 report but using three factors; they reported Cronbach's alphas of .77-.94. de Kruijff, McWilliam, Ridley, and Wakely (2000) referred to Arnett (1989) and the original four factors, but chose to focus on only two dimensions—sensitivity and detachment—and to combine these subscales together into a single overall score "for ease of interpretation" (pp. 253).

The theory that motivated the development of the Arnett CIS implies that the CIS should correlate with child outcomes; however, as is true for other child care quality measures (Burchinal, Kainz, & Cai, 2011), past studies found that associations between the CIS and child developmental outcomes are typically small. These studies have employed the Arnett CIS in various ways including with a single total score (Hindman, Skibbe, Miller, & Zimmerman, 2010; Lisonbee, Mize, Payne, & Granger, 2008; Loeb, Fuller, Kagan, & Carrol, 2004; Zill et al., 2003), with factor scores (Howes, Galinsky, & Kontos, 1998; Vernon-Feagans & Manlove, 2005; Whitebook et al., 1989), and as part of a composite variable created along with other measures of child care quality (Burchinal & Cryer, 2003; Peisner-Feinberg et al., 2001; Votruba-Drzal, Coley, & Chase-Lansdale, 2004; Votruba-

Drzal et al., 2010). Specifically, using a composite measure including the Arnett CIS and three other observational measures of quality, Peisner-Feinberg et al. (2001) found small contemporaneous associations with child outcomes including child language, reading and math scores (effect sizes of .06 to .18) and with child attention, problem behaviors, and sociability (effects sizes of .03 to .11). The four studies cited above that used the Arnett CIS total score generally found no significant associations with children's cognitive and socio-emotional outcomes, although Loeb et al. (2004) found a modest association with children's greater reading readiness ( $r = .14$ ) and fewer social problems ( $r = -.18$ ). The three studies that associated Arnett CIS subscales with child outcomes found significant associations with attachment, complex play, and verbalizations. Whitebook et al. (1989) found small partial correlations (.24 and below), Howes et al. (1998) did not report effect sizes, & Vernon-Feagans and Manlove (2005) reported large effect sizes, although their findings were specific to nine children with chronic ear infections who were cared for in low-quality contexts.

The tendency for prior studies to create composite measures of child care quality from the Arnett CIS and other scales reflects the fact that quality measures tend to be highly inter-correlated (Burchinal & Nelson, 2000). For example, Burchinal and Cryer (2003) reported correlations between .74 and .91 among the Arnett CIS, the Early Childhood Environmental Environment Rating Scale-Revised (ECERS-R) and another quality measure. Indeed, studies such as those listed above often pair the Arnett CIS with the ECERS-R, regarding the latter as a broader measure of the caregiver environment and the former as a narrower measure of caregiver-child interactions. Yet, subscales and items within the ECERS-R also assess interactions among the caregiver and children, which may contribute to the high inter-correlation between the two scales.

## 1.2. The current study

The Arnett CIS has been widely adopted in studies of child care with limited published evidence of its validity. The fact that there have been few replications of Arnett's original factor structure, and the absence of consistently significant and sizable associations between Arnett CIS and child developmental outcomes suggests that more research into the psychometric properties of the Arnett CIS is warranted. Our study contributes to the literature by examining the factor structure of the Arnett CIS in a national sample of toddlers and preschoolers, by conducting item response theory (IRT) analyses of the items, and by examining regression-adjusted associations of the resulting dimensions with child outcomes. To the best of our knowledge, our study is the first to report IRT analyses of the Arnett CIS. We included toddlers even though Arnett developed the measure for a study of child care teachers caring for preschoolers (Arnett, 1986, 1989), because the studies reviewed used the scale for infants and toddlers and in home-based settings.

We framed our analyses and results within four aspects of validity included in the Standards for Educational and Psychological Testing: structural validity, response process validity, test content validity, and criterion validity (Joint Committee on Standards for Educational and Psychological Testing, 1999).

**1.2.1. Structural validity: number of dimensions**—A measure is structurally valid if empirical evidence identifies its intended dimensions (Wolfe & Smith, 2007). Arnett did not report an explicit *a priori* definition of dimensions for the scale. Instead, he focused on percent agreement between raters in developing the items and use of empirical techniques to define the scale's dimensions (Arnett, 1986, 1989). We test for his four empirical dimensions (or other structures found by later work, such as the first three factors or a single factor). As we discuss in Section 2, the fact that the Arnett CIS mixes positively oriented

and negatively oriented items, and that the orientation of items is nearly completely confounded with its empirically identified dimensions, complicates analysis of the scale's dimensional structure.

**1.2.2. Response process validity: ordering of rating scale categories—**A measure demonstrates response process validity if observers use its rating scale categories in intended ways (Joint Committee on Standards for Educational and Psychological Testing, 1999). For the Arnett CIS, observers assign a score of 1 = Not at all, 2 = Somewhat, 3 = Quite a bit, or 4 = Very much for each item. The labels imply that caregivers placed in higher categories on positively oriented items (e.g. speaks warmly to the children) should have higher positions on the latent construct of positive interactions with children. We used adjacent-category models that allowed us to test for this expected category order.

**1.2.3. Test content validity: item quality—**Test content validity refers to the relationship between a set of items and the latent trait that they are intended to measure (Joint Committee on Standards for Educational and Psychological Testing, 1999). If this aspect of validity holds, the set of items defining each of the Arnett CIS' four factors (see again Table 1) should consistently measure single underlying dimensions. The items in each dimension should also evaluate a wide range of difficulty within the dimension (i.e., some easy items, some mid-range items, some difficult items) in order to ensure sufficiently large variation (given that the sample is heterogenous) and minimal standard errors for resulting scale scores (given that the items are well targeted to the sample). As noted above, Arnett did not use a contemporary measurement approach to define dimensions and did not explicitly write items to cover the full range of those dimensions. Thus, our post hoc analyses point to potential gaps in coverage of the dimensions which future measurement work might address.

**1.2.4. Criterion validity: correlations with child outcomes—**Tests of criterion validity determine the extent to which instrument scores correlate with relevant outcome measures (Wolfe & Smith, 2007). Arnett based the CIS in the literature on parenting behavior as suggested by Baumrind's (1967) classification—authoritative, authoritarian, permissive, and disengaged. Without familiarity with this literature, it may seem at first blush that the Arnett CIS should primarily promote children's socio-emotional development. However, the theory and the empirical evidence suggest cross-domain associations (Downer, Sabol, & Hamre, 2010). Indeed, parenting types have been correlated with a range of child outcomes, not only the absence of negative affect and aggression but also child complexity of play and academic achievement (Brown & Iyengar, 2008; Lagacé-Séguin & d'Entremont, 2006; Underwood, Beron, Gentsch, Galperin, & Risser, 2008), since parental and caregiver sensitivity provide children a secure base to explore their environment and the confidence to try new experiences (Denham & Burton, 2003). Caregiver involvement and monitoring also may support child health, such as through caregivers' interventions when injury risk is high and vigilance to consistent infection control practices (like child handwashing).

## 2. Method

### 2.1. Data

We used data from the Early Childhood Longitudinal Study-Birth Cohort (ECLS-B). The study began in 2001 when children were sampled from birth records in 46 states. The ECLS-B oversampled twins, low birth weight babies, American Indian, and Asian children. The sample size at the initial 9-month interview was 10,700, resulting from a 74% response rate.

At the 2- and 4-year interviews, 9850 and 8950 children remained, respectively (Snow et al., 2009).

We focused on the subgroup of 2- and 4-year-olds whose child care settings were observed. At both waves, statisticians drew a stratified random sample of children who met the following criteria: in child care for at least 10 h per week (and awake for at least a 2.5 h block of time) who also lived in the continental U.S. and whose caregiver spoke English or Spanish. At 2 years, the ECLS-B included all poor children (less than 100% of the federal poverty level) cared for in centers and a subsample of the remaining children (with differential sampling of twins, low-birth-weight, American Indian and Asian children). At 4 years, the ECLS-B included all children whose care centers were observed at 2 years and sampled the remaining children based on type of care (home-based care, Head Start centers, and non-Head Start centers) and child poverty (less than 100%, 100–150%, and over 150% of the federal poverty level; Wheelless, Ault, & Park, 2008). Response rates of selected child care providers were 50% at 2 years and 55% at 4 years (Bethel et al., 2007; Wheelless et al., 2008). The ECLS-B statisticians created sampling weights that adjusted for oversampling, family non-response initially and over time, and caregiver non-response (W22P0 at 2 years; W33P0 at 4 years; Snow et al., 2007). We focused on observed home-based providers at 2 years ( $n = 750$ ) and observed center-based providers at 4 years ( $n = 1350$ ). We did so because these are the modal care arrangements at each follow-up, offering adequate samples for analyses. In regression models, with item-level missing data, the sample sizes were 650 at 2 years and 1000 at 4 years.

## 2.2. Measures

We provide descriptive statistics for all measures in Appendix A (included in online supplementary material).

**2.2.1. Arnett CIS**—Trained observers completed the Arnett CIS in homes and classrooms including the ECLS-B focal child (Snow et al., 2007). ECLS-B initially certified observers who achieved 80% agreement with a consensus score and monitored reliability by having two observers rate approximately 10% of settings. Percent agreement averaged 96–98% over the field periods (Nord, Edwards, Andreassen, Green, & Wallner-Allen, 2006; Snow et al., 2009). The items were worded as in the rightmost column of Table 1 and were rated 1 = Not at all, 2 = Somewhat, 3 = Quite a bit, or 4 = Very much. Table 2 provides the percentage of the 750 homes and 1350 centers that received each scale score (1 to 4) on each item. Before calculating the percentage distributions, we reverse scored negatively worded items (shown with an (R) at the end of their descriptive label in the table). We also grouped items by modal category in centers and highlighted the modal category in centers and homes. The distributions indicated that the items were quite skewed with most caregivers demonstrating the positive behaviors captured by each item. The lowest two scores were never modal for centers and were modal for only three items for homes. For centers, at least half of the caregivers were given one of the two highest scores on every item; for homes, this was true for all but two items. And, on fully half of the items in both types of care, 50% or more of caregivers were rated in the category showing the most positive behavior: *Very much*. We used these Arnett CIS items in IRT models.

**2.2.2. Parenting and caregiving measures**—Given the Arnett CIS was based on a well-established theory of parenting, we compare the strength of associations between the Arnett CIS and child outcomes with the strength of associations between an observational measure of parenting and child outcomes. Likewise, since the child care literature often compares structural quality to process quality and the parenting literature often compares family structure to family process, we associate structural measures of the family and

childcare context with child outcomes as a point of comparison. Estimates of associations between these measures, the Arnett CIS, and child outcomes were obtained using regression analyses.

The process parenting measure was the Two Bags Task, a standardized, semi-structured observational measure of parenting adapted for the ECLS-B from earlier versions of the measure (Three Bags Task and Puzzle Task; Chase-Lansdale, Brooks-Gunn, & Zamsky, 1994; Fauth, Brady-Smith, & Brooks-Gunn, 2003; Matas, Arend, & Sroufe, 1978). Trained ECLS-B interviewers followed standard procedures to videotape interactions between the same focal children that were observed with the Arnett CIS and one of their parents. The parent (typically the mother) and child were observed playing with toys from “two bags” for 10 min in the family’s home. The contents of the bags were a play set of dishes and picture book at 2 years and molding clay and a book at 4 years. Trained observers had to achieve target standards of percent agreement on initial training videotapes and on reliability checks during the field period. Average percent agreement was 87–100% across parenting items (Nord et al., 2006; Snow et al., 2007).

At 2 years, observers coded six parenting items on a 7-point Likert Scale: positive regard, sensitivity, stimulation of cognitive development, intrusiveness, negative regard, and detachment. At 4 years, positive regard and sensitivity were combined into a single item: emotional supportiveness. In line with previous research that has found two broad constructs of parenting styles (Barnett, Deng, Mills-Koonce, Willoughby, & Cox, 2008; Pungello, Iruka, Dotterer, Mills-Koonce, & Reznick, 2009), we combined the items into two domains: (1) cognitive stimulating/sensitive and (2) harsh/intrusive. We calculated the first parenting domain as the mean of positive regard, sensitivity, stimulation of cognitive development and detachment (reversed) at 2 years and the mean of emotional supportiveness, stimulation of cognitive development and detachment (reversed) at 4 years. We averaged intrusiveness and negative regard to calculate the second parenting domain at both waves.

At both the 2-year and 4-year waves, ECLS-B staff created a family socioeconomic status composite by averaging five variables each of which was first standardized to a mean of 0 and standard deviation of 1: (1) father/male guardian’s education, (2) mother/female guardian’s education, (3) father/male guardian’s occupational prestige score, (4) mother/female guardian’s occupational prestige score, and (5) household income. For households with only one parent, the ECLS-B staff averaged the three available components (resident parent’s education and occupation as well as household income; Snow et al., 2007).

We also examined three measures of the structural quality of the child care setting: (1) the teacher or caregiver’s level of education, (2) the number of children in the classroom/home, and (3) the ratio of children to teachers/caregivers in the classroom/home. The caregiver/teacher reported all of these measures. Group size refers to the number of children that the teacher “typically” cared for at the same time as the focal child and the child:caregiver ratio was based on the group size along with the number of adults who “usually” helped care for that group. Teachers and caregivers were instructed to include their own children, if relevant, in the group size.

**2.2.3. Child outcome measures**—To facilitate interpretation across numerous measures, we coded all child outcomes so that higher scores indicate better outcomes.

**2.2.4. Cognitive outcome measures**—To measure cognitive outcomes in the 2-year-old sample, we used a mental score created by the ECLS-B staff from the Bayley Short Form-Research Edition (BSF-R; Andreassen & Fletcher, 2007), which was adapted for the ECLS-B from the Bayley Scales of Infant Development-Second Edition (BSID-II; Bayley,

1993). The scale assessed the children's emerging cognitive skills such as putting objects in a container, imitating words, naming pictures, and attending to stories (Andreassen & Fletcher, 2007; Snow et al., 2007, 2009). ECLS-B staff used IRT methods to produce a model-based estimate of the BSF-R mental scale raw score, which can range from 92 to 174 (Nord et al., 2006).

For 4-year-olds, we used two scores that ECLS-B staff created in the domains of reading and math using item response theory models (Najarian, Snow, Lennon, Kinsey, & Mulligan, 2010; Snow et al., 2007, 2009). The ECLS-B investigators selected items from three subtests of the Preschool Language Assessment Scales (Simon Says, Art Show and Let's Tell Stories; Duncan & De Avila, 1998), from the Peabody Picture Vocabulary Test-Third Edition (Dunn & Dunn, 1997), and a measure of emergent early literacy including letter sounds, early reading, phonological awareness, knowledge of print conventions, and matching words (Snow et al., 2007, 2009). For the mathematics assessment, ECLS-B investigators took items from the Test of Early Mathematics Ability, the ECLS-K (Early Childhood Longitudinal Study-Kindergarten Cohort survey; Snow et al., 2007), and other sources in the following areas: number sense, geometry, counting, operations, and patterns (Najarian et al., 2010; Snow et al., 2007, 2009). ECLS-B reported IRT-based internal consistency reliability estimates of .84 for reading and .89 for math (Najarian et al., 2010).

**2.2.5. Socio-emotional outcomes**—At 2 years, we created socio-emotional composite scores based on interviewer-rated items that the ECLS-B study designers selected from the BSID's Behavior Rating Scale for the BSF-R in three areas: (1) social competence (e.g., child displays social engagement, child displays cooperation), (2) emotional and behavioral regulation (e.g., child displays frustrations [reversed], fearfulness [reversed], positive affect), and (3) attention and concentration (child pays attention, child is persistent in tasks, child adapts to change in material; Andreassen & Fletcher, 2007; Nord et al., 2006). We averaged the items, which ranged from 1 to 5. Cronbach's alpha values were good to excellent (.87 for social competence, .93 for emotional and behavioral regulation, .96 for attention and concentration). We also used a composite measure of the child's temperament based on seven items selected by ECLS-B staff from the Infant/Toddler Symptom Checklist (ITSC; DeGangi, Poisson, Sickel, & Weiner, 1995; Nord et al., 2006). We summed the seven items, which were all rated on a scale from 0 to 3 and captured whether the child was frequently fussy or irritable, easily went from a whimper to a cry, was unable to wait without crying, was easily distractible, needed help to fall asleep, tuned out from activities, and could not shift focus easily.

At 4 years, the ECLS-B study designers selected items from the Preschool and Kindergarten Behavior Scales-Second Edition (PKBS-2; Merrell, 2003), Social Skills Rating System (SSRS; Gresham & Elliott, 1990), and ECLS-K which we organized into three constructs: (1) social competence (e.g., how well the child plays with others, is liked by others, and is accepted by others), (2) emotional and behavioral regulation (e.g., lack of aggression, anger, and worry; expressions of happiness), and (3) attention and concentration (child pays attention well, does not disrupt the class, and is not overly active). Both parents and caregivers/teachers rated these preschool socio-emotional items. We averaged the items, which were scored from 1 to 5. Cronbach's alpha values were acceptable to excellent (for parents and caregivers/teachers respectively: .83 and .97 for social competence, .68 and .94 for emotional and behavioral regulation, .69 and .94 for attention and concentration).

**2.2.6. Health outcomes**—Although prior developmental research has less often considered health outcomes, we expect that caregiver involvement and monitoring may, for example, promote vigilance to child safety (thus reducing injuries) and to hygiene (thus reducing illnesses). At both 2 and 4 years, parents rated children's health on a score of 1



(excellent) to 5 (poor). Most children were rated as healthy; thus we created a dichotomous indicator: *excellent* (coded 1), or *very good, good, fair, or poor* (coded as 0). The parent also reported whether the child had a doctor-verified *respiratory illness, gastrointestinal illness, or ear infection*, and whether the child had experienced an *injury* that required a doctor's visit since the last interview. We created individual dummy variables to indicate the absence of illness or injury and a sum of the three indicators of absence of illness. At 2 years, we also used the BSF-R motor scale to measure the children's acquisition of motor skills, such as reaching and grasping, manipulating small objects, and balancing and walking (Andreassen & Fletcher, 2007; Snow et al., 2007,2009). ECLS-B staff used IRT methods to produce a model-based estimate of the BSF-R motor scale raw score, which can range from 21 to 87 (Nord et al., 2006).

**2.2.7. Control Variables**—In analyses of criterion validity, we adjusted for a number of covariates that may be associated with both the quality of child care and with child outcomes.

**2.2.8. Child-level controls**—Child covariates at both 2 years and 4 years included: child gender and racial and Hispanic identification (dummy coded as Hispanic, non-Hispanic Black, or non-Hispanic of other race versus non-Hispanic White), whether the child was born low birth weight, whether the child was ever breast fed, whether the child had two or more well-child doctor visits since the last interview, and whether the child had received WIC since the last interview.

We also controlled for lagged child outcomes at both waves. For the 4-year sample these included the measures of cognitive, socioemotional and health outcomes at 2 years, described above. For the 2-year sample, we created similar 9-month dummy indicators of mother rating of child excellent health, sum of absence of illness, and the ITSC measure of the child's temperament.

**2.2.9. Family level controls**—Family demographic covariates included: whether the mother was born outside of the U.S., whether there were any other children less than age 6 or any children ages 6 to 18 in the household, the mother's marital status, the mother's employment status, maternal age, whether the family had used TANF and whether the family had used Food Stamps since the last interview.

**2.2.10. Community-level controls**—To adjust for cross-region variation, we included dummy indicators for region of residence (South, Midwest, West, or Northeast), and urbanicity of the ZIP Code (rural, urban area of fewer than 50,000 people, urban area of 50,000 people or more).

**2.2.11. Child care-level controls**—We coded the center teacher or home-based provider's gender, age, race-ethnicity, experience (years), and certification in early childhood education (at 2 years, a dummy indicator of whether the caregiver had a certificate in early childhood education, other areas of education, nursing, social work, or psychology; at 4 years, the sum of five indicators of whether the teacher had various coursework or credentials in early childhood education). We also controlled for the length of time (months) the child had been at the observed child care setting and the hours per week that the child currently attended that setting. For centers, we also created variables to measure the location and funding stream (Head Start, public school, private school, religious school/church, community non-profit or community for-profit), accreditation status, whether the center accepted subsidies, licensing status, and, if licensed, the number of children for which the center was licensed to provide care.

### 3. Analytic approach

We relied on classical test theory (e.g., correlation, factor analysis, and regression) and item response theory approaches to provide evidence on the four aspects of construct validity outlined above.

#### 3.1. Structural validity: number of dimensions

We examined structural validity with a series of weighted and unweighted confirmatory factor analyses and unweighted IRT approaches. Because we found similar results across specifications (see Appendix B, included in online supplementary material), we present our unweighted IRT models—a series of multidimensional generalized partial credit models (GPCM) estimated using IRTPRO statistical software (Version 2.1; Cai, Thissen, & du Toit, 2011). The GPCM is appropriate for multi-category rating scales such as the Arnett CIS that have multiple subscales (Embretson & Reise, 2000). We examined traditionally structured GPCM models as well as bifactor GPCM models. The bifactor model is estimated similarly to traditional models, but imposes a particular structure on the data. The two components implied by the “bi” prefix are: (1) a general factor on which every item loads and (2) a set of additional factors wherein each item loads on only one additional factor (Gibbons et al., 2007; Gibbons & Hedeker, 1992; Holzinger & Swineford, 1937). The set of additional factors are uncorrelated with one another and with the general factor. The bifactor model has been used to distinguish general from specific substantive factors (e.g., a general extraversion personality trait and specific aspects of extraversion such as warmth, gregariousness, and assertiveness; Chen, Hayes, Carver, Laurenceau, & Zhang, 2012) and to separate a general substantive factor from specific methodological factors (a general reading comprehension skill measured with items nested within paragraphs that vary in content; Gibbons & Hedeker, 1992). We use the bifactor model for the latter purpose, to separate a substantive factor from a set of methodological factors. Specifically, we anticipated that some of the correlation among the Arnett CIS items was due to their positive or negative orientation. That is, raters may respond in a somewhat different way to the set of positively oriented items than to the set of the negatively oriented items (i.e., the way each observer thinks of the response category “very much” may differ somewhat when he or she rates a positively oriented item as compared to when he or she rates a negatively oriented item). The bifactor model washes away this correlation due to item orientation. The general factor is then a better measure of the substantive dimension, in our case caregiver–child interactions.

Because prior studies had sometimes identified more than one substantive dimension in the Arnett CIS, we extended the bifactor approach by allowing for multiple substantive factors that were uncorrelated with the set of methods (“item orientation”) factors. We based the substantive dimensions on Table 1 including all four of the dimensions listed in the table, as well as one-, two-, and three-dimensional models (based on the highest crossdimension correlations, the three-dimensional model combined Permissiveness and Harshness; the two-dimensional model combined Sensitivity and Detachment as well as Permissiveness and Harshness). As noted, the methodological factors reflected the orientation of the items (see Table 2, where negatively oriented items are indicated by an [R]). For example, in the four-dimensional model, we placed Arnett18 on the Permissiveness substantive dimension and the positively oriented methodological dimension; we also placed Arnett15 on the Permissiveness substantive dimension but the negatively oriented methodological dimension (see again Table 2).

Our attempt to identify multiple substantive factors along with the two methods factors was unsuccessful, however. That is, the bifactor models with two-, three-, or four-substantive dimensions failed to converge. This is because the substantive and methodological

dimensions are nearly completely confounded in the Arnett CIS. Again, examining Table 2 it is clear that all of the Sensitivity items are positively oriented; all but two of the Permissive, Harshness and Detachment items are negatively oriented (one item each from the Permissiveness and Harshness dimensions are positively oriented); and none of the Detachment items are positively oriented. The traditionally structured (substantive only) four-dimensional model also failed to converge. Thus, we report results only for the uni-dimensional bifactor GPCM model and the uni-, two- and three-dimensional traditionally structured GPCM models.

We relied on the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) to choose among models. A difference in AIC and BIC values of 10 or more provides strong evidence for the model with the smaller value (Burnham & Anderson, 2002; Raftery, 1995). We also calculated deviance statistics (the log-likelihood multiplied by  $-2$ ). The difference in deviance statistics between nested models is distributed as a chi-square with degrees of freedom equal to the difference in the number of parameters between the two models. A statistically significant chi-square test indicates that the model with the smaller deviance (i.e., larger log-likelihood) is preferred.

**3.1.1. Response process validity: ordering of categories**—We also used the GPCM to examine the ordering of the rating scale categories. The GPCM allows for this analysis because it does not force order between adjacent categories (Andrich, 1996; Andrich, de Jong, & Sheridan, 1997). We used 95% confidence intervals for the thresholds that separate adjacent categories to define *order*, *overlap*, and *disorder*. That is, within items, we defined two adjacent categories' thresholds as: (a) *ordered* if the upper bound of the confidence interval for the lower threshold was below the lower bound of the confidence interval for the next higher threshold, (b) *overlapping* if the confidence intervals of the lower threshold and higher adjacent threshold overlapped, and (c) *disordered* if the upper bound of the confidence interval for next higher adjacent threshold was below the lower bound of the confidence interval for the adjacent lower threshold. Because IRTPRO parameterizes the link functions so that monotonically decreasing thresholds are associated with an increasing amount of the latent trait, we multiplied the thresholds by negative one before applying these criteria.

### 3.2. Test content validity: item quality

We also used the GPCM to calculate two item statistics commonly reported in psychometric IRT studies: item difficulty (decomposed into an overall difficulty and category thresholds) and discrimination estimates (Embretson & Reise, 2000). Overall item difficulty is an average difficulty level that locates the item along the caregiver interaction spectrum. Estimates are on a logit scale, ranging between about  $-3$  and  $3$ , with lower (negative) values indicating easier items (which means that most caregivers are more likely to be rated with the higher categories) and higher (positive) values indicating harder items (which means that most caregivers are less likely to be rated with the higher categories). Discrimination estimates represent the degree to which an item can distinguish between caregivers of higher and lower quality, indicating the amount of information each item provides.

### 3.3. Criterion validity: associations with caregiver characteristics and with measures of child outcomes

The GPCM analyses also produced estimates of each dimension. In a bifactor model like ours that distinguishes a general substantive factor from a set of methodological factors, only the substantive factor is of interest for assessing criterion validity. We thus produce an estimate of the caregiver's positive interactions, which we refer to as the *Arnett CIS Measure*. As indicated in Tables 2, 4 and 5, negatively oriented items were reverse scored

prior to model estimation, so higher scores reflect more positive interactions. These quality measures are on the same scale as the item difficulty estimates (in logit units). We first examined the associations between these Arnett CIS Measures and caregiver characteristics. For continuous characteristics, we calculated Pearson correlations. For categorical characteristics, we calculated t-tests. In both cases, we applied the ECLS-B sampling weight.

We also used the IRT measures in regression analyses to examine criterion validity. For continuous outcomes, we estimated OLS regression models and standardized the coefficients by multiplying the unstandardized coefficient by the standard deviation of the focal predictor and dividing by the standard deviation of the outcome. For dichotomous outcomes, we estimated logit regression models and reported discrete change, which is the difference in the predicted probability when the focal predictor is one half of a standard deviation above the mean versus one half of a standard deviation below the mean with all other predictors held constant at their means (Gordon, 2012; Long, 1997). As benchmarks for comparing the size of the coefficients, we used parenting, family SES, and child care structural quality as predictors in similar regression models. We also explored interactions by including a dummy variable to indicate whether or not the child had been in the observed child care setting both for more than 6 months and for more than 15 h per week. We explored this interaction because we anticipated that associations with the Arnett CIS Measure might be larger when the child had more exposure to the caregiver. In regression models, we controlled for the covariates in Appendix A (included in online supplementary material) and adjusted for oversampling and non-response with the ECLS-B weights.

We additionally estimated models in which we either logged quality or added a squared term for quality, in order to test for possible non-linear associations. In all but two cases, the R-squared values for these alternative models rounded to the same value as the comparable linear model (within two decimal places) and the squared terms were non-significant. The exception was two models predicting parent reports of preschoolers' social competence and attention/concentration, where the squared terms were significant. Predicted values revealed that in these cases, parents rated children best (as most attentive and most socially competent) when the caregiver showed the lowest and highest scores on the Arnett CIS. Since these results were unanticipated, and were not replicated for caregiver reports (nor at 2 years), we viewed them as tentative and focused on the linear estimates below.

## 4. Results

### 4.1. Sample description

Before turning to our tests of validity, we describe our two samples (Appendix A provides full descriptive statistics, included in online supplementary material). Doing so highlights some of the differences between caregivers of 2-year-olds in homes and teachers of 4-year-olds in centers, which are expected given the distinct regulations and typical organizations of center-based and home-based care. These differences should be kept in mind when interpreting our findings. Beginning with structural quality characteristics, teachers of 4-year-olds averaged over 3.5 more years of education than did caregivers of 2-year-olds, which is consistent with differences in licensing and accreditation standards. Again reflecting different standards, preschool teachers also cared for children in larger groups (almost 11 more children, on average) and with larger ratios (almost three more children per adult, on average). It is also the case that preschool teachers averaged over 3 more years in the child care field than caregivers of 2-year-olds (although the standard deviation of years of experience was also almost 10 times larger for the home-based caregivers). Children also spent more time in home-based settings at 2 years than in center-based contexts at 4 years. The toddlers averaged nearly 40 h per week in the observed child care homes, and had been

there for over a year, on average. In contrast, the preschoolers averaged just over 20 h per week in the observed child care centers, and had been there for just over 6 months, on average.

#### 4.2. Structural validity (Number of dimensions)

Table 3 presents model fit indices for the series of GPCM. For both homes and centers, the bifactor model fit best. Within each type of child care, the deviance, AIC and BIC values were smallest for the bifactor model (bolded values in Table 3). In all cases, the differences in fit were sizable: differences in AIC and BIC values between the bifactor model and the traditionally structured GPCM models (with only substantive dimensions) were all substantially larger than 10; and, the chi-square values for the differences in deviance values between nested models indicated statistical significance at  $p < .05$ . These results confirm the importance of adjusting empirically for the shared method variance among positively and negatively oriented items, and suggest that the factors that prior scholars interpreted substantively may reflect this method variance instead (although as we discuss further below, the structure of the Arnett CIS makes it difficult to definitively separate substantive and methodological dimensions).

#### 4.3. Response process validity (Ordering of categories)

We found little evidence of disorder in the category threshold estimates. In the sample of homes, none of the items had any disordered category thresholds (results not shown). In the sample of centers, just one item had a disordered category threshold (Arnett13: Spends time not involved). More items demonstrated overlapping category thresholds, but about half the items (14 of 26 items for centers, 12 of 26 items for homes) had completely ordered thresholds. Among those with some overlap, most items in centers showed overlap in only one of the two adjacent threshold comparisons (10 of 12 items), specifically between the lowest two category thresholds (i.e., the category thresholds separating 1 versus 2 and 2 versus 3 overlapped), which generally corresponded to the least used categories. For home-based care, about half of the items (6 of 14 items) had just one overlapping threshold, again primarily between the lowest two category thresholds. For both homes and centers, the one item with disorder and the remaining items with overlapping regions were the negatively oriented items (with (R) notations in Table 2). No positively oriented items had disorder or overlap.

The minimal disorder suggests that raters are generally using the categories as expected (higher scores reflect higher quality), especially for positively oriented items. The overlapping thresholds evident for the negatively oriented items likely reflects their skewness (see again Table 2); very few cases in the lowest categories increase standard errors which contribute to overlap.

#### 4.4. Test content validity (item quality)

Tables 4 and 5 present the item difficulty and discrimination estimates for homes and centers, respectively. Results show that all discrimination values were well above zero for the substantive dimension, as well as, for the most part, the methodological dimensions. In fact, the supermajority was above 1.0, which suggests that the items are generally informative. Nearly all item difficulty levels were negative, however, reflecting the skewness that we saw in Table 2. That is, most Arnett CIS items were easy—they reflect characteristics often demonstrated by most teachers and caregivers in the ECLS-B.

#### 4.5. Criterion validity (Associations with caregiver/teacher characteristics and with measures of child outcomes)

Table 6 shows the correlations of the Arnett CIS Measures with caregiver/teacher characteristics. The top panel shows results for categorical characteristics, providing the within-subgroup average of the Arnett CIS Measure and indicating significant differences in means between subgroups with shared subscripts. The bottom panel shows results for continuous characteristics, providing the Pearson correlation with the Arnett CIS Measure and indicating statistical significance with an asterisk.

Numerous caregiver characteristics were associated with the Arnett CIS Measure, but not always in the same way for homes and centers. The most consistent finding was that more educated home-based caregivers and preschool teachers were rated higher on the Arnett CIS Measure, with correlations of .20 to .24. Certification was also associated with better ratings on the Arnett CIS. Based on the categorical measure for home-based caregivers, those with some certification in early childhood education or a related field scored .22 points higher (.41 –.19), which was almost one-third of a standard deviation ( $SD = .78$ ). Based on the continuous measure of certification for center-based teachers, we saw a significant correlation, although small at .11. Race-ethnicity was also associated with Arnett CIS Measures, but in different ways for the two types of care. Among caregivers of 2-year-olds in homes, non-Hispanic whites scored higher on the Arnett CIS, on average, than did non-Hispanic Blacks or Hispanics. Among teachers of 4-year-olds in centers, non-Hispanic Blacks scored lower on the Arnett CIS, on average, than each other racial-ethnic group. We also saw that age was positively associated with the Arnett CIS in centers, and group size was negatively associated with the Arnett CIS in homes (both small in size). Male teachers also averaged higher scores than did female teachers in centers (although the sample size of men was small, at just 3% of the approximately 1000 cases).

Tables 7 and 8 show the standardized associations of the Arnett CIS Measure with child outcomes, adjusting for the child, family, community and child care controls in Appendix A (included in online supplementary material). For comparison, we also provide associations with child outcomes for child care structural measures and for family process and structural measures.

Beginning with Table 7, we see that the Arnett CIS Measure was significantly positively associated with four child outcomes among 2-year-olds in home-based care: the BSF-R mental score and all three socio-emotional outcomes (social, regulation, and attention). Associations were small in size with standardized coefficients ranging from .11 to .15. Still, associations with Arnett compare favorably to associations with family process and structure, especially for socio-emotional development where the Two Bags Task Stimulating/Sensitive Score had a significant association of .17 in size, the Two Bags Task Harsh/Intrusive Score had a significant association of  $-.09$  in size, and Family SES had significant associations of .17 to .21 in size. The associations of family process and structure with 2-year-olds emerging cognition were also small in size (.10 to .26 in magnitude), although Family SES and Stimulating/Sensitive parenting had about double the association of the Arnett CIS Measure.

The Arnett CIS Measure was not associated with any health outcomes for 2-year-olds in home-based care. The absence of illness was associated with other constructs, however. Two-year-olds were less likely to be illness-free when their home-based provider cared for more children and when their family scored higher on Stimulating/Sensitive parenting and SES (all small associations, ranging from .13 to .15 in magnitude).

In Table 8, we see that the Arnett CIS Measure is generally not associated with outcomes for 4-year-olds in center care including none of the cognitive and socio-emotional outcomes and all but one health outcome. The exception is that children were more likely to be free from injuries when their center teacher scored higher on the Arnett CIS (although the association was small, at .06). Most other associations with child care structure and family process/structure were also small. The largest associations were between Family SES and 4-year-olds' cognitive development (.28 for math and .29 for reading). The Two Bags Task Stimulating/Sensitive measure of parenting was also positively associated with preschoolers' reading scores, although the magnitude was small (.10). In these center-based contexts, we also saw that preschoolers' math scores were somewhat higher, on average, in larger settings, with more children per caregiver. Again the association was small, at .08.

Finally, we explored interactions between the Arnett CIS Measure and dummy indicators representing whether the focal child had spent more months and more hours per week in the observed child care setting. These results are important for understanding the differences in results we see for homes and centers because the toddlers in our sample spent substantially more time with their observed home-based caregiver than the preschoolers spent with their teachers. Indeed, the two samples were at opposite poles of exposure. That is, fully one-fifth (21%) of preschoolers spent 15 or fewer hours per week in the observed center and had been there for 6 months or less. In contrast, almost one-third (29%) of toddlers spent 40 or more hours per week in the observed home-based care setting and had been there for at least 18 months. Viewed another way, the majority (over three quarters) of preschoolers spent less than 40 h per week in the center and had been there for less than 18 months whereas the majority (fully 75%) of toddlers spent more than 15 h per week in the home and had been there for at least 6 months.

In our regression models, we found that exposure did not moderate the associations between the Arnett CIS Measure and child outcomes for 2-year-olds in homes, perhaps not surprising given that most had high exposure. In contrast, exposure did moderate the associations for the cognitive and some caregiver-reported socio-emotional outcomes of 4-year-olds in centers. Specifically, we found that among the 300 preschoolers who had been with the teacher for more than 6 months and for more than 15 h per week, the Arnett CIS Measure was significantly associated with higher math and reading scores ( $\beta = .16, p < .04$ ;  $\beta = .14, p < .05$ ). For the remaining 750 preschoolers who had been with the teacher for 6 months or less or for 15 h per week or less, the associations were not significant ( $\beta = -.09, p < .10$  and  $\beta = -.02, p < .62$  for math and reading, respectively).

The difference in slopes in the two groups was statistically significant for math and approached significance for reading ( $p < .01$  and  $p < .06$  respectively). Similarly, among the preschoolers with high exposure, the Arnett CIS Measure was significantly associated with higher caregiver-reported emotional and behavioral regulation scores ( $\beta = .16, p < .03$ ) and higher attention and concentration scores ( $\beta = .11, p < .05$ ). The difference in slopes between the two groups was also statistically significant for emotional and behavioral regulation and approached significance for attention and concentration ( $p < .05$  and  $p < .09$ , respectively). We did not find evidence of moderation for parent-reported socio-emotional nor for health outcomes among 4-year-olds.

## 5. Discussion

Our study adds to the handful of psychometric analyses of the Arnett CIS, and as far as we are aware, is the first to apply item response theory (IRT) models. We do not find evidence for the four dimensions identified by Arnett's original principal component analysis, nor the two or three dimensions evident in some prior studies. Instead, we find that a bifactor model

fits best, which reveals a single substantive dimension (caregiver interaction) while controlling for variance that arises due to a pair of methodological dimensions (item orientation). The results suggest that factors that have sometimes been interpreted as separate substantive dimensions in prior studies (e.g., sensitive versus detached caregiver-child interactions) may be better viewed as method variance. Our adjacent-category IRT models also did not force item categories to be ordered, allowing us to account for the few categories that were disordered and the several that were overlapping. This improves upon prior studies that have relied on exploratory or confirmatory factor analyses which assumed item categories followed an ordinal progression.

Our IRT models also verified what is evident in descriptive item tabulations: the items capture behaviors that most caregivers in this national sample display. In other words, there is little variation among the items. In psychometric terms, this means that the items are not well targeted at the sample. Better targeting would include items that capture behaviors that only few and some, but not most, caregivers display. This poor targeting of the items on the sample is not something that the bifactor model can correct, and the resulting scale is thus less informative than it would be if the items were better targeted. We find that regression-adjusted associations with child outcomes are sometimes significant both for toddlers in home-based care and for preschoolers in centers. However, among preschoolers, the associations are only significant for children in centers for the longest hours and most months; most toddlers in our sample have this more extensive exposure of at least 6 months and at least 15 h per week in their home care settings. Although small in absolute terms, we found significant associations for cognitive and socio-emotional outcomes, as expected, and the magnitudes of these associations were similar to associations between an observational measure of parenting and child outcomes. Thus, the criterion validity of the scale is promising, although better targeting of the items would further improve its validity.

Our study has some limitations. Although the ECLS-B has a nationally representative sampling design, just half of the sampled providers agreed to be observed. It is possible that the distribution of items would be less skewed in another sample. As far as we are aware, however, the ECLS-B is the most representative sample of Arnett scores collected to date. Because of its scope, the ECLS-B also rarely implemented standardized measures of children's cognitive and socio-emotional development in their entirety and instead created short forms of existing measures or drew items from multiple sources. The scale developers were often involved in creating these short forms and ECLS-B psychometricians did considerable empirical verification of many of the scales. However, it is possible that associations would be larger between the Arnett CIS and full forms of standardized measures of child development. It is also possible that other measures of child development such as children's stress, moods, or engagement, would be more highly associated with caregiver sensitivity (Dunn & Kontos, 1997).

With these limitations in mind, our findings have important implications for use of the Arnett CIS and for scale development. The early childhood field is challenged by the generally small associations between child care quality and child outcomes such as those we find in our study. On the one hand, these findings may reflect the concept of "good enough" parenting, which views development as robust, adaptive, and to a large extent genetically determined, except at the neglectful or abusive extreme of parenting behavior (Scarr, 1992, 1996). On the other hand, critics of "good enough" parenting suggest that this viewpoint overstates the importance of heredity and understates the importance of parenting, especially for vulnerable children (Baumrind, 1993; Jackson, 1993). Indeed, early intervention studies suggest that intensive preschool interventions can improve lifetime socioeconomic trajectories in cost-effective ways, especially for poor, minority children (Heckman & Masterov, 2007; Webster-Stratton & Taylor, 2001). A challenge for the early childhood field



is to distinguish between two competing explanations of the modest associations between child care quality and developmental outcomes: that preschool and child care quality are not measured well because they have not undergone the intensive measurement scrutiny that has typified measures of other developmental constructs (e.g., DeRoos & Allen-Meaers, 1998; Piquero, Macintosh, & Hickman, 2002; Rapport, LaFond, & Sivo, 2009), or that child care quality does not have a true, substantial association with developmental outcomes. Our research demonstrates some specific avenues for measurement improvement, including that the Arnett CIS includes method variance that may have clouded associations in prior studies and that its skewed items limit variation in the measure and, as a consequence, associations with developmental outcomes.

As noted above, although the bifactor model corrected the issue of method variance, it could not correct for the limited variation. That is, scores on the Arnett CIS are clustered on the positive end, with most caregivers/teacher demonstrating the rated behaviors. With the paucity of items that fewer caregivers demonstrate, we have little information to distinguish among the most skilled caregivers. On the one hand, it is not surprising that most caregivers are sensitive. Preschool teachers' training and the professional norms of the field promote sensitivity and family day care providers select into the field because they enjoy being with children (Blau, 1997; Burchinal, Howes, & Kontos, 2002; Fukkink & Lont, 2007). Other measures of teacher sensitivity and warmth show similar skew. For example, the CLASS authors' found that Emotional Support averaged between 5 and 6 on a 7-point scale over the course of a school year, as did Classroom Organization, but Instructional Support averaged only between 2 and 3 (Pianta, La Paro, & Hamre, 2008). Still, if a purpose of measuring caregiver sensitivity is to predict child outcomes, then it is important to have ample variation. We anticipate that achieving such variation for the Arnett CIS would require revising existing items, writing additional items, and modifying the response structure to better differentiate between "moderately" and "highly" sensitive caregivers.

Our review of the content of the items more specifically suggests that they often cover relevant constructs, but that they sometimes require subjective inference, pack in multiple constructs (especially with parenthetical examples), and use subjective response labels. For example, Item 21 "Fails to show interest in the children's activities" includes three parenthetical examples: removes self from children's activities, doesn't talk to children, and [doesn't] extend [children's] conversations. The examples have a natural hierarchy, in which more sensitive caregivers might move from being present during the activity, to talking with children about the activity, to extending children's conversations during the activity. By packing these together in a single item, observers may vary in the criteria they use to determine whether the caregiver "fails to show interest" (must the caregiver demonstrate none of the activities across the entire observation period to be rated as "not at all"?). Separating the criteria into three separate items might allow for better identification of caregivers who demonstrate none, one, two or all three of the behaviors.

Making the response structure less subjective, and more frequency based, might also improve variation. In the example just discussed, observers could rate each of the three behaviors in terms of the number or fraction of time that they are observed. The current response structure may lead observers to be reluctant to rate caregivers outside of the "Not at all" category on negatively oriented items because the wording and scale imply a general trait of the caregiver. For example, scoring outside of the "Not at all" category on Item 2 "Seems critical of the children (e.g., puts children down, uses sarcasm)," might be seen as a statement that "This is a caregiver who is critical of children." In contrast, if observers rated on a frequency scale how many times the caregiver put a child down or used sarcasm during the observation period, they might be more likely to give scores outside of the "None" category (e.g., "This caregiver showed four instances of being critical of children"). The

latter approach allows observers to rate the caregiver with fewer subjective interpretations and also rate the “state” of the caregiver’s behavior in specific situations (rather than suggesting the caregiver has a “trait” of being generally critical of children). Future methodological studies might explore whether such alternative frequency or percentage type scoring approaches captures more variation in caregiver negative behavior than the original Arnett CIS category labels.

As we noted above, the Arnett CIS was grounded in a well-established theory of parenting, but an iterative psychometric scale development process was not followed. Arnett did not explicitly define dimensions nor write items to capture various levels of those dimensions. Thus, an important starting point for revisions would be a better definition of the dimensions, beginning with Baumrind’s original conceptualization of parenting styles but extending to more recent conceptualizations. Indeed, Baumrind’s approach was explicitly person-centered rather than variable-centered (Baumrind, 2013); and, other scholars have since contributed to the elaboration of the two fundamental dimensions often seen as underlying her four typologies: warmth/responsiveness and control/discipline (Maccoby & Martin, 1983; Morris et al., 2013). Particularly important in contemporary thinking is the distinction between coercive or psychological control, which impede healthy development, and confrontive or behavioral control, which promote healthy development; authoritative parents demonstrate the latter but not the former (Barber & Xia, 2013; Baumrind, 2013). Contemporary research is identifying the ways in which children can perceive strategies such as time-out or positive reinforcement as power assertive, leading to less healthy emotional and behavioral regulation (Bergin & Bergin, 1999; Denham & Burton, 2003; Readdick & Chapman, 2000). Extending such current thinking about parenting to caregiving in family day care and preschool contexts might be an important place to begin in updating the Arnett CIS.

Indeed, the lack of specific connections to concepts underlying Baumrind’s typology is evident when scrutinizing the Arnett CIS items, and may explain why we find evidence of a single substantive dimension. For example, the items in the ‘Harshness’ and ‘Permissiveness’ dimensions often mix together rigidity of schedule (coercive control) with a harsh tone (lack of warmth) suggesting the absence of a clear, single, defining construct. The fact that recent users of the Arnett CIS added parenthetical examples to the original item wording also reflects the subjectivity and ambiguity of the items (e.g., what is “too strongly?” what is “unnecessarily harsh?”), consistent with the general finding in the survey methodology field that examples are often added in an attempt to clarify a complex concept when rewriting a question for more focus and clarity is a better solution (Schaeffer & Presser, 2003). The Arnett CIS items also implicitly reflect a professional definition of quality that has been critiqued for its singular cultural lens (Cryer, 1999). Observers that endorse more didactic teaching or more authoritarian parenting styles might rate the scale differently than those reflecting the child-initiated and authoritative approach embedded in the scale, and the scale might be a less valid measure of quality for children and caregivers reflecting the former cultural perspective (Cryer, 1999; Shivers, Sanders, & Westbrook, 2011). Methodological studies that randomized observers to receive different versions of the scale (with and without the examples; with newly worded items) could illuminate these issues of ambiguity, subjectivity, and cultural relevance.

Although we found the bifactor model fit well, identifying a single substantive caregiving factor and adjusting for a pair of methodological factors, it would also be desirable in future scale development to have numerous items capturing each substantive dimension and to not confound the orientation of items with substantive dimensions. Arnett’s original third and fourth dimensions had just 3–4 items, which provide less information for estimating with precision the underlying latent factor than the other two dimensions, which had 9–10 items.

It is also the case that the items for Arnett's first dimension (sensitivity) were positively oriented while nearly all other items were negatively oriented. In future development work, it would be preferable to positively orient all items, or to balance positively and negatively oriented items across substantive dimensions (Stone, 2004).

A recent dissertation offers a first attempt at modifying the Arnett CIS, resulting in a new measure called the Child Caregiver Interaction Scale (CCIS; Carl, 2007, 2010). Although the CCIS is yet unpublished, Carl reinforces the issues documented in our study, also critiquing the Arnett CIS' items and response structure for their subjectivity, highlighting the problems with the skewness of the items, and noting the lack of definitions of dimensions and items. She aimed to update the scale's framing using constructs of developmentally appropriate practice and strategies of recent measures, including reorganizing the items into three subscales in emotional, cognitive, and social domains (more consistent with the CLASS) and using an indicator/item structure (following the ECERS). However, she reports only exploratory factor analyses and her three dimensions are highly correlated (.75 to .87). She did not report IRT methods, which may be a concern since recent research identified problems with the ECERS-R response structure (Gordon, Fujimoto, Kaestner, Korenman, & Abner, 2012). She did develop the CCIS for different age groups and for homes and centers. In fact, one of the reasons the Arnett CIS has found such widespread use may be that it is not as age- or context-dependent as other measures such as the Environment Rating Scales, which have different versions by type of care and child age (Harms, Clifford, & Cryer, 1998), or the CLASS, which was originally developed explicitly for center-based preschool classrooms (Pianta et al., 2008).

In conclusion, our analyses suggest that the Arnett CIS measures one substantive dimension rather than four subscales. Although the items could be improved to better measure this dimension and most of the items reflect behaviors that most caregivers do, correlations with child outcomes compare favorably to correlations with family background and parenting, especially for toddlers in homebased care and for preschoolers with at least a half year of at least half-time exposure to centers. We recommend that future users of the Arnett CIS test for the bifactor structure; and, if verified, use its single substantive dimension in analyses. We also encourage further attempts to improve the measure and future empirical study of the reliability and validity of new measures, such as those discussed above.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A090065, and by the Eunice Kennedy Shriver National Institute of Child Health and Human Development, through Grant R01HD060711. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education. We thank Everett Smith for his expert consultation on our psychometric analyses and Clancy Blair, Jeanne Brooks-Gunn, Anna Johnson, and Lauren Wakschlag for their expert consultation on the relevance of Arnett items for supporting various domains of child development.

## Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.ecresq.2012.12.004>.

## References

- Andreassen, C.; Fletcher, P. Early childhood longitudinal study birth cohort (ECLS-B): Psychometric report for the 2-year data collection (NCES 2007-084). Washington, DC: National Center for Education Statistics; 2007.
- Andrich, D. Measurement criteria for choosing among models with graded responses. In: Eye, AV.; Clogg, CC., editors. *Categorical variables in developmental research: Methods of analysis*. San Diego, CA: Academic Press; 1996. p. 3-35.
- Andrich, D.; de Jong, JH.; Sheridan, BE. Diagnostic opportunities with the Rasch model for ordered response categories. In: Rost, J.; Langeheine, R., editors. *Applications of latent trait and latent class models in the social sciences*. New York, NY: Waxmann Verlag GMBH; 1997. p. 58-68.
- Arnett, J. Caregivers in day care centers: Does training matter?. Charlottesville, VA: University of Virginia; 1986. (Unpublished doctoral dissertation)
- Arnett J. Caregivers in day-care centers: Does training matter? *Journal of Applied Developmental Psychology*. 1989; 10(4):541–552. [http://dx.doi.org/10.1016/0193-3973\(89\)90026-9](http://dx.doi.org/10.1016/0193-3973(89)90026-9).
- Barber, BK.; Xia, M. The centrality of control to parenting and its effects. In: Larzelere, RE.; Morris, AS.; Harrist, AW., editors. *Authoritative parenting: Synthesizing nurturance and discipline for optimal child development*. Washington, DC: American Psychological Association Press; 2013. p. 61-88.
- Barnett MA, Deng M, Mills-Koonce WR, Willoughby M, Cox M. Interdependence of parenting of mothers and fathers of infants. *Journal of Family Psychology*. 2008; 22(4):561–573. <http://dx.doi.org/10.1037/0893-3200.22.3.561>. [PubMed: 18729670]
- Baumrind D. Child-care practices anteceding three patterns of preschool behavior. *Genetic Psychology Monographs*. 1967; 75:43–88. [PubMed: 6032134]
- Baumrind D. The average expectable environment is not good enough: A response to Scarr. *Child Development*. 1993; 64:1299–1317. [PubMed: 7693400]
- Baumrind, D. Authoritative parenting revisited: History and current status. In: Larzelere, RE.; Morris, AS.; Harrist, AW., editors. *Authoritative parenting: Synthesizing nurturance and discipline for optimal child development*. Washington, DC: American Psychological Association Press; 2013. p. 11-34.
- Bayley, N. Bayley scales of infant development. 2nd ed.. San Antonio, TX: The Psychological Corporation; 1993.
- Bergin C, Bergin DA. Classroom discipline that promotes self-control. *Journal of Applied Developmental Psychology*. 1999; 20:189–206.
- Bethel, J.; Green, JL.; Kalton, G.; Nord, C.; Mulligan, GM.; Eyster, S. Early childhood longitudinal study, birth cohort (ECLS-B), sampling report for the 2-year data collection. Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U. S. Department of Education; 2007.
- Blau DM. The production of quality in child care centers. *The Journal of Human Resources*. 1997; 32:354–387.
- Bracken SS, Fischel JE. Assessment of preschool classroom practices: Application of q-sort methodology. *Early Childhood Research Quarterly*. 2006; 21:417–430.
- Brown L, Iyengar S. Parenting styles: The impact on student achievement. *Marriage & Family Review*. 2008; 43(1):14–38.
- Burchinal MR, Cryer D. Diversity, child care quality, and developmental outcomes. *Early Childhood Research Quarterly*. 2003; 18:401–426.
- Burchinal M, Howes C, Kontos S. Structural predictors of child care quality in child care homes. *Early Childhood Research Quarterly*. 2002; 17:87–105.
- Burchinal, M.; Kainz, K.; Cai, Y. How well do our measures of quality predict child outcomes? A meta-analysis and coordinated analysis of data from largescale studies of early childhood settings. In: Zaslow, M.; Martinez-Beck, I.; Tout, K.; Halle, T., editors. *Quality measurement in early childhood settings*. Baltimore, MD: Brookes; 2011. p. 11-31.
- Burchinal M, Nelson L. Family selection and child care experiences: Implications for studies of child outcomes. *Early Childhood Research Quarterly*. 2000; 15:385–411.

- Burnham, KP.; Anderson, DR. Model selection and multimodel inference: A practical information-theoretic approach. 2nd ed.. New York, NY: Springer; 2002.
- Cai, L.; Thissen, D.; du Toit, S. IRTPRO2.1 [computer software and manual]. Skokie, IL: Scientific Software International; 2011.
- Carl, B. Child caregiver interaction scale. Indiana, PA: Indiana University of Pennsylvania; 2007. (Unpublished doctoral dissertation)
- Carl B. Child caregiver interaction scale (CCIS) revised edition manual. 2010 Unpublished instrument. Retrieved from author.
- Chase-Lansdale PL, Brooks-Gunn J, Zamsky ES. Young African-American multigenerational families in poverty: Quality of mothering and grandmothering. *Child Development*. 1994; 65:373–393. [PubMed: 8013228]
- Chen FF, Hayes A, Carver CS, Laurenceau J, Zhang Z. Modeling general and specific variance in multifaceted constructs: A comparison of the bifactor model to other approaches. *Journal of Personality*. 2012; 80:219–251. [PubMed: 22092195]
- Cryer D. Defining and assessing early childhood program quality. *Annals of the American Academy of Political and Social Science*. 1999; 563(1):39–55.
- Cryer D, Tietz W, Burchinal M, Leal T, Palacios J. Predicting process quality from structural quality in preschool programs: A cross-country comparison. *Early Childhood Research Quarterly*. 1999; 14:339–361.
- DeGangi, DA.; Poisson, S.; Sickel, RZ.; Weiner, AS. Infant/toddler symptom checklist: a screening tool for parents. San Antonio, TX: The Psychological Corporation; 1995.
- Denham, SA.; Burton, R. Social and emotional prevention and intervention programming for preschoolers. New York, NY: Kluwer; 2003.
- de Kruif REL, McWilliam RA, Ridley SM, Wakely MB. Classification of teachers' interaction behaviors in early childhood classrooms. *Early Childhood Research Quarterly*. 2000; 15:247–268.
- DeRoos Y, Allen-Meares P. Application of Rasch analysis: Exploring differences in depression between African-American and White children. *Journal of Social Service Research*. 1998; 23:93–107.
- Downer J, Sabol TJ, Hamre B. Teacher-child interactions in the classroom: Toward a theory of within- and cross-domain links to children's developmental outcomes. *Early Education and Development*. 2010; 21:699–723.
- Duncan, SE.; De Avila, EA. PreLAS 2000. Monterey, CA: CTB/McGraw-Hill; 1998.
- Dunn, LM.; Dunn, LM. Peabody picture vocabulary test-third edition (PPVTIII). Upper Saddle River, NJ: Pearson; 1997.
- Dunn L, Kontos S. What have we learned about developmentally appropriate practice? *Young Children*. 1997; 52:4–13.
- Embretson, SE.; Reise, SP. Item response theory for psychologists. Mahwah, NJ: Erlbaum; 2000.
- Fauth, RC.; Brady-Smith, C.; Brooks-Gunn, J. Parent-child interaction rating scales for the Play Doh® task and father-child interaction rating scales for the three-bag task. New York, NY: National Center for Children and Families (NCCF), Teachers College, Columbia University; 2003.
- Fukkink RG, Lont A. Does training matter? A meta-analysis and review of caregiver training studies. *Early Childhood Research Quarterly*. 2007; 22:294–311.
- Gibbons RD, Bock RD, Hedeker D, Weiss DJ, Segawa E, Bhaumik DK, et al. Full-information item bifactor analysis of graded response data. *Applied Psychological Measurement*. 2007; 31(1):4–19.
- Gibbons RD, Hedeker D. Full-information item bifactor analysis. *Psy-chometrika*. 1992; 57(3):423–436.
- Gordon, RA. Applied statistics for the social and health sciences. New York, NY: Routledge; 2012.
- Gordon, RA.; Fujimoto, K.; Kaestner, R.; Korenman, S.; Abner, K. An assessment of the validity of the ECERS-R with implications for measures of child care quality and relations to child development. *Developmental Psychology*. Advance online publication. 2012. <http://dx.doi.org/10.1037/a0027899>
- Gresham, FM.; Elliott, SN. Social skills rating system manual. Circle Pines, MN: American Guidance Service; 1990.

- Harms, T.; Clifford, RM.; Cryer, D. Early childhood environment rating scale revised. New York, NY: Teachers College Press; 1998.
- Heckman JJ, Masterov DV. The productivity argument for investing in young children. *Review of Agricultural Economics*. 2007; 29:446–493.
- Hindman AH, Skibbe LE, Miller A, Zimmerman M. Ecological contexts and early learning: Contributions of child, family, and classroom factors during Head Start, to literacy and mathematics growth through first grade. *Early Childhood Research Quarterly*. 2010; 25:235–250.
- Holzinger KJ, Swineford F. The bifactor method. *Psychometrika*. 1937; 2(1):41–54.
- Howes C, Galinsky E, Kontos S. Child care caregiver sensitivity and attachment. *Social Development*. 1998; 7:25–36.
- Jackson JF. Human behavioral genetics Scarr’s theory, and her views on interventions: A critical review and commentary on their implications for African American children. *Child Development*. 1993; 64:1318–1332. [PubMed: 7693401]
- Joint Committee on Standards for Educational and Psychological Testing. Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 1999.
- Layzer, JI.; Goodson, BD.; Moss, M. Observational study of early childhood programs: Life in preschool final report. Cambridge, MA: Abt Associates; 1993.
- Lagacé-Séguin DG, d’Entremont M-R. The role of child negative affect in the relations between parenting styles and non-adaptive play. *Early Child Development and Care*. 2006; 176(5):461–477.
- Lisonbee JA, Mize J, Payne AL, Granger DA. Children’s cortisol and the quality of teacher-child relationships in child care. *Child Development*. 2008; 79:1818–1832. [PubMed: 19037952]
- Loeb S, Fuller B, Kagan SL, Carrol B. Child care in poor communities: Early learning effects of type, quality and stability. *Child Development*. 2004; 75:47–65. [PubMed: 15015674]
- Long, JS. Regression models for categorical and limited dependent variables. Thousand Oaks, CA: Sage; 1997.
- Love, J.; Constantine, J.; Paulsell, D.; Boller, K.; Ross, C.; Raikes, H., et al. The role of Early Head Start programs in addressing the child care needs of low-income families with infants and toddlers: Influences on child care use and quality. Washington, DC: U. S. Administration for Children and Families; 2004.
- Love, J.; Ryer, P.; Faddis, B. Caring environments-program quality in California’s publicly funded child development programs: Report on the legislatively mandated 1990-91 staff/child ratio study. Portsmouth, NH: RMC Research Corporation; 1992.
- Maccoby, EE.; Martin, JA. Socialization in the context of the family: Parent-child interaction. In: Mussen, PH.; Hetherington, EM., editors. *Handbook of child psychology: Socialization, personality, and social development*. 4th ed. Vol. 4. New York, NY: Wiley; 1983. p. 1-101.
- Matas L, Arend RA, Sroufe LA. Continuity and adaptation in the second year: The relationship between quality of attachment and later competence. *Child Development*. 1978; 49:547–566.
- Maxwell KL, McWilliam RA, Hemmeter ML, Ault MJ, Schuster JW. Predictors of developmentally appropriate classroom practices in kindergarten through third grade. *Early Childhood Research Quarterly*. 2001; 16:431–452.
- McCartney K, Scarr S, Phillips D, Grajek S. Day care as intervention: Comparisons of varying quality programs. *Journal of Applied Developmental Psychology*. 1985; 6:247–260.
- Merrell, KM. Preschool and kindergarten behavior scales (PKBS-2). Austin, TX: Pro-Ed; 2003.
- Mollborn S, Blalock C. Consequences of teen parents’ child-care arrangements for mothers and children. *Journal of Marriage and Family*. 2012; 74:846–865. <http://dx.doi.org/10.1111/j.1741-3737.2012.00988.x>. [PubMed: 23729861]
- Morris, AS.; Cui, L.; Steinberg, L. Parenting research and themes: What we’ve learned and where to go next. In: Larzelere, RE.; Morris, AS.; Harrist, AW., editors. *Authoritative parenting: Synthesizing nurturance and discipline for optimal child development*. Washington, DC: American Psychological Association Press; 2013. p. 35-57.
- Najarian, M.; Snow, K.; Lennon, J.; Kinsey, S.; Mulligan, G. Early childhood longitudinal study, birth cohort (ECLS-B): Preschool-kindergarten 2007 psychometric report (NCES2010-009). Washington, DC: National Center for Education Statistics; 2010.

- Nord, C.; Edwards, B.; Andreassen, C.; Green, JL.; Wallner-Allen, K. Early childhood longitudinal study-birth cohort (ECLS-B): user's manual for the ECLS-B longitudinal 9-month-2-year data file and electronic codebook (NCES 2006-046). Washington, DC: National Center for Education Statistics; 2006.
- Peisner-Feinberg ES, Burchinal MR, Clifford RM, Culkin ML, Howes C, Kagan SL, et al. The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child Development*. 2001; 72:1534–1553. [PubMed: 11699686]
- Pianta, RC.; La Paro, KM.; Hamre, BK. Classroom assessment scoring system -PreK manual. Baltimore, MD: Brookes; 2008.
- Piquero AR, Macintosh R, Hickman M. The validity of a self-reported delinquency scale: Comparisons across gender, age, race, and place of residence. *Sociological Methods and Research*. 2002; 30:492–529.
- Pungello EP, Iruka IU, Dotterer AM, Mills-Koonce R, Reznick JS. The effects of socioeconomic status, race, and parenting on language development in early childhood. *Developmental Psychology*. 2009; 45(2):544–557. [PubMed: 19271838]
- Raftery AE. Bayesian model selection in social research. *Sociological Methodology*. 1995; 25:111–163.
- Rapport MD, LaFond SV, Sivo SA. One-dimensionality and developmental trajectory of aggressive behavior in clinically-referred boys: A Rasch analysis. *Journal of Psychopathology and Behavioral Assessment*. 2009; 31:309–319.
- Readdick CA, Chapman PL. Young children's perceptions of time out. *Journal of Research in Childhood Education*. 2000; 15:81–87.
- Scarr S. Developmental theories for the 1990s: Development and individual differences. *Child Development*. 1992; 63:1–19. [PubMed: 1343618]
- Scarr S. How people make their own environments: Implications for parents and policy makers. *Psychology, Public Policy, and Law*. 1996; 2:204–228.
- Scarr S, McCartney K. Far from home: An experimental evaluation of the mother-child home program in Bermuda. *Child Development*. 1988; 59:531–543.
- Schaeffer NC, Presser S. The science of asking questions. *Annual Review of Sociology*. 2003; 29:65–88.
- Shivers, EM.; Sanders, K.; Westbrook, TR. Measuring culturally responsive early care and education. In: Zaslow, M.; Martinez-Beck, I.; Tout, K.; Halle, T., editors. *Quality measurement in early childhood settings*. Baltimore, MD: Brookes; 2011. p. 191-225.
- Snow, K.; Derecho, A.; Wheelless, S.; Lennon, J.; Kinsey, S.; Morgan, K., et al. Early childhood longitudinal study, birth cohort (ECLS-B): Kindergarten 2006 and 2007 data file user's manual (NCES 2010-010). Washington, DC: National Center for Education Statistics; 2009.
- Snow, K.; Thalji, L.; Derecho, A.; Wheelless, S.; Lennon, J.; Kinsey, S., et al. Early childhood longitudinal study, birth cohort (ECLS-B): Preschool year data file user's manual (2005–06) (NCES 2008–024). Washington, DC: U. S. Department of Education, Institute of Education Sciences, National Center for Education Statistics; 2007.
- Stone, MH. Substantive scale construction. In: Smith, EV., Jr.; Smith, RM., editors. *Introduction to Rasch measurement*. Maple Grove, MN: JAM; 2004. p. 201-225.
- Tietz W, Cryer D, Bairrao J, Palacios J, Wetzel G. Comparisons of observed process quality in early child care and education programs in five countries. *Early Childhood Research Quarterly*. 1996; 11:447–475.
- Torquati JC, Raikes H, Huddleston-Casas CA. Teacher education, motivation, compensation, workplace support, and links to quality of center-based child care and teachers' intention to stay in the early childhood profession. *Early Childhood Research Quarterly*. 2007; 22:261–275.
- Underwood MK, Beron KJ, Gentsch JK, Galperin MB, Risser SD. Interparental conflict resolution strategies, parenting styles, and children's social and physical aggression with peers. *International Journal of Behavioral Development*. 2008; 32:566–579.
- van IJzendoorn MH, Tavecchio LWC, Stams GJJM, Verhoeven MJE, Reiling EJ. Quality of center day care and attunement between parents and caregivers: Center day care in cross-national

- perspective. *The Journal of Genetic Psychology: Research and Theory on Human Development*. 1998; 159(4):437–454. <http://dx.doi.org/10.1080/00221329809596163>.
- Vernon-Feagans L, Manlove EE. Otitis media, the quality of child care, and the social/communicative behavior of toddlers: A replication and extension. *Early Childhood Research Quarterly*. 2005; 20:306–328.
- Votruba-Drzal E, Coley RL, Chase-Lansdale PL. Child care and low-income children's development: Direct and moderated effects. *Child Development*. 2004; 75:296–312. [PubMed: 15015691]
- Votruba-Drzal E, Coley RL, Maldonado-Carreo C, Li-Grining CP, Chase-Lansdale PL. Child care and the development of behavior problems among economically disadvantaged children in middle childhood. *Child Development*. 2010; 81:1460–1474. [PubMed: 20840234]
- Webster-Stratton C, Taylor T. Nipping early risk factors in the bud: Preventing substance abuse, delinquency and violence in adolescence through interventions targeted at young children (0–8 years). *Prevention Science*. 2001; 2:165–192. [PubMed: 11678292]
- Wheeless, S.; Ault, K.; Park, J. Early childhood longitudinal study birth cohort (ECLS-B): Methodology report for the preschool data collection (2005–06) sampling. Washington, DC: National Center for Education Statistics; 2008.
- Whitebook, M.; Howes, C.; Phillips, D. Who cares: Child care teachers and the quality of care in America. Oakland, CA: Child Care Employee Project; 1989.
- Wolfe, EW.; Smith, EV, Jr.. Instrument development tools and activities for measure validation using Rasch models: Part II - validation activities. In: Smith, EV., Jr.; Smith, RM., editors. *Rasch measurement: Advanced and specialized applications*. Maple Grove, MN: Journal of Applied Measurement Press; 2007. p. 243-290.
- Zill, N.; Resnick, G.; Kim, K.; O'Donnell, K.; Sorongon, A.; McKey, R., et al. Head start FACES 2000 A whole-child perspective on program performance (HHS-105-96-1912). Washington, DC: Administration for Children and Families, Child Outcomes Research and Evaluation; 2003.
- Zill, N.; Resnick, G.; Kim, K.; O'Donnell, K.; Sorongon, A.; Ziv, Y., et al. Head start performance measures center family and children experiences survey (FACES 2000): Technical report [Executive Summary]. Washington, DC: U.S. Administration for Children and Families, Office of Planning, Research and Evaluation; 2006.



**Table 1**

Dimensions and item wordings on the Arnett CIS scale.

<b>Dimension</b>	<b>Item</b>	<b>Original wording of the Arnett CIS</b>	<b>Wording in the ECLS-B</b>
Sensitivity	1	Speaks warmly to the children	Speaks warmly to the children (e.g., positive tone of voice, body language)
	3	Listens attentively when children speak to her	Listens attentively when children speak to her (e.g., looks at children, nods, rephrases their comments, engages in conversations)
	6	Seems to enjoy the children	Seems to enjoy the children (e.g., conveys warmth by smiling, touching, taking children's conversations seriously)
	7	When children misbehave, explains the reason for the rule they are breaking	When children misbehave, explains the reason for the rule they are breaking (e.g., discusses consequences, redirects behavior, discusses what to do instead)
	8	Encourages the children to try new experiences	Encourages the children to try new experiences (e.g., suggests children do it together, helps children start, introduces new materials)
	11	Seems enthusiastic about the children's activities and efforts	Seems enthusiastic about the children's activities and efforts (e.g., congratulates children, states appreciation for their efforts)
	14	Pays positive attention to the children as individuals	Pays positive attention to the children as individuals (e.g., speaks to individual children, uses their names, calls attention to prosocial behaviors, comments on their strengths)
	16	Talks to the children on a level they can understand	Talks to the children on a level they can understand (e.g., uses terms familiar to children, checks for clarification)
	19	Encourages children to exhibit prosocial behavior	Encourages children to exhibit prosocial behavior (e.g., sharing, cooperating, pairs socially skillful with those children that need practice)
	25	When talking to children, kneels, bends, or sits at their level to establish better eye contact	When talking to children, kneels, bends or sits at their level to establish better eye contact (e.g., ensures connection when having a conversation)
Harshness	2	Seems critical of the children	Seems critical of the children (e.g., puts children down, uses sarcasm)
	4	Places high value on obedience	Places high value on obedience (e.g., expects children to follow adult agenda, fails to respond to daily events in a flexible)
	10	Speaks with irritation or hostility to the children	Speaks with irritation or hostility to the children (e.g., sharp tone, raises voice)
	12	Threatens children in trying to control them	Threatens children in trying to control them (e.g., uses bribes and threats of punishment)
	17	Punishes the children without explanation	Punishes the children without explanation (e.g., does not discuss infraction)
	20	Finds fault easily with the children	Finds fault easily with the children (e.g., negative tone, critical)
	22	Seems to prohibit many of the things the children want to do	Seems to prohibit many of the things the children want to do (e.g., adheres to rigid schedule or adult outcomes and agendas)
	24 <sup>a</sup>	Expects the children to exercise self-control, e.g., to be undistruptive for group, teacher-led activities, to be able to stand in line calmly	Expects the children to exercise a reasonable amount of self-control (e.g., expects children to be undistruptive for short group, teacher-led activities; to be able to stand in line calmly; reminds children of expectations; and asks for cooperation in supportive ways)
26	Seems unnecessarily harsh when scolding or	Seems unnecessarily harsh when scolding or prohibiting children (e.g., angry)	

Dimension	Item	Original wording of the Arnett CIS	Wording in the ECLS-B
		prohibiting children	tone, shakes children, uses physical punishment, uses "time-out" without explanation)
Detachment	5	Seems distant or detached from the children	Seems distant or detached from the children (e.g., sits apart, does not touch children, does not greet children)
	13	Spends considerable time in activity not involving interaction with the children	Spends considerable time in activity not involving interaction with the children (e.g., does adult tasks during child activity periods)
	21	Doesn't seem interested in the children's activities	Fails to show interest in the children's activities (e.g., removes self from children's activities, doesn't talk to children or extend their conversations)
	23	Doesn't supervise the children very closely	Fails to supervise the children very closely (e.g., withdraws during activities, fails to foresee and forestall mishaps)
Permissiveness	9	Doesn't try to exercise much control over the children	Exercises too much control over the children (e.g., doesn't take child input, rigid adherence to rules and schedules)
	15	Doesn't reprimand children when they misbehave	Reprimands children too strongly when they misbehave (e.g., is punitive, fails to acknowledge difficulties of learning self-control, fails to redirect behavior)
	18	Exercises firmness when necessary.	Exercises firmness when necessary (e.g., clear and direct directions, checks for understanding)

*Source.* Original wording: Arnett (1986). Wording in the ECLS-B: Snow et al. (2007).

<sup>a</sup>Arnett (1986) placed Item 24 on the fourth (Permissiveness) dimension. We placed it on the second (Harshness) dimension where the ECLS-B placed it.

Table 2

Percentage of settings rated at each scale value for each Arnett CIS item, sorted by modal category in centers.

Item	Subscale	Homes (2-year sample)				Centers (4-year sample)			
		1 = Not at all	2 = Somewhat	3 = Quite a Bit	4 = Very much	1 = Not at all	2 = Somewhat	3 = Quite a Bit	4 = Very much
Amnett7: When misbehave explains reason	Sensitivity	12%	41%	31%	15%	4%	29%	39%	28%
Amnett19: Encourages prosocial behavior	Sensitivity	6%	32%	39%	23%	2%	16%	41%	40%
Amnett25: Kneels bends or sits at child level	Sensitivity	7%	28%	39%	25%	3%	21%	41%	35%
Amnett8: Encourages new experiences	Sensitivity	12%	42%	32%	15%	5%	23%	43%	29%
Amnett18: Exercises firmness	Permissiveness	6%	36%	45%	13%	2%	16%	48%	34%
Amnett24: Expects self-control	Harshness	10%	39%	37%	14%	4%	15%	48%	34%
Amnett11: Seems enthusiastic about activity	Sensitivity	6%	30%	36%	28%	3%	18%	38%	41%
Amnett3: Listens attentively	Sensitivity	1%	21%	42%	36%	1%	14%	40%	44%
Amnett14: Pays positive attention to children	Sensitivity	3%	25%	40%	33%	2%	17%	33%	48%
Amnett16: Talks on level children can understand	Sensitivity	1%	18%	48%	34%	1%	8%	41%	49%
Amnett6: Seems to enjoy the children	Sensitivity	1%	17%	39%	43%	2%	14%	33%	51%
Amnett1: Speaks warmly to children	Sensitivity	0%	12%	40%	49%	1%	12%	33%	54%
Amnett4: Places high value on obedience (R)	Harshness	4%	9%	23%	64%	5%	11%	26%	58%
Amnett22: Prohibits many things (R)	Harshness	1%	2%	15%	82%	2%	5%	22%	70%
Amnett9: Exercise too much control (R)	Permissiveness	1%	3%	11%	85%	3%	6%	18%	73%
Amnett13: Spends time not involved (R)	Detachment	3%	8%	27%	63%	2%	3%	18%	77%
Amnett21: Doesn't show interest (R)	Detachment	2%	5%	21%	72%	1%	3%	17%	79%
Amnett23: Doesn't supervise closely (R)	Detachment	1%	5%	23%	70%	1%	3%	16%	80%
Amnett2: Threatens control (R)	Harshness	1%	4%	17%	78%	2%	3%	14%	81%
Amnett5: Seems distant or detached (R)	Detachment	1%	3%	18%	78%	1%	2%	15%	82%
Amnett10: Speaks with irritation (R)	Harshness	1%	2%	13%	84%	1%	3%	14%	82%
Amnett15: Reprimands too strongly (R)	Permissiveness	1%	2%	10%	86%	2%	4%	12%	83%
Amnett20: Finds fault easily (R)	Harsh	1%	2%	10%	87%	2%	3%	9%	86%
Amnett2: Critical of children (R)	Harsh	0%	1%	9%	90%	1%	2%	10%	87%
Amnett17: Punishes without explanation (R)	Harsh	0%	1%	5%	93%	1%	2%	10%	88%
Amnett26: Unnecessarily harsh (R)	Harsh	1%	1%	7%	91%	1%	3%	8%	88%

*Note.*  $n = 750$  at 2 years.  $n = 1350$  at 4 years. The modal category for each item is shaded. The items are ordered first by the modal category at 4 years and second by the percentage in that modal category. Negatively worded items were reverse coded before calculating percentage distributions, as indicated by an “(R)” at the end of the item description (e.g., Arnett4 is reverse coded).

**Table 3**

Measures of fit from one-dimensional bifactor and the one-, two- and three-dimensional traditionally structured (substantive only) generalized partial credit model.

	Homes (2-year sample)			Centers (4-year sample)						
	Traditional (substantive only) GPCMs			Traditional (substantive only) GPCMs						
	Bifactor	One dimension	Two dimensions (original)	Two dimensions (modified) <sup>a</sup>	Three dimensions <sup>b</sup>	Bifactor	One dimension	Two dimensions (original)	Two dimensions (modified) <sup>a</sup>	Three dimensions <sup>b</sup>
Deviance	<b>25375.5</b>	26302.7	26045.2	25837.5	26185.1	<b>44136.1</b>	45408.2	45034.5	44674.5	45445.7
AIC	<b>25635.5</b>	26510.7	26255.2	26047.5	26353.1	<b>44396.1</b>	45616.2	45244.5	44884.5	45613.7
BIC	<b>26234.2</b>	26989.6	26738.7	26531.1	26789.5	<b>45071.5</b>	46156.5	45790.0	45430.0	46050.1

Note.  $n = 750$  at 2 years.  $n = 1350$  at 4 years. Bolded values are the smallest values within each row and sample. Dimensional structures based on conceptual layout of items in Table 1. The three-dimensional model combines Permissiveness and Harshness in one dimension, the two-dimensional combines Permissiveness and Harshness for the first dimension and Sensitivity and Detachment for the second dimension. The four-dimensional structure shown in Table 1 failed to converge, and thus no results are shown for it.

<sup>a</sup>Items 18 and 24 placed on the Sensitivity/Detachment combined dimension.

<sup>b</sup>The three dimensional structure constrained the discrimination parameters to be the same value across the items within a dimension.

**Table 4**

Overall item difficulty and discrimination estimates for homes (2-year sample) based on the bifactor generalized partial credit model.

Item	Substantive dimension		Method dimensions	
	Difficulty	Discrimination	Positively phrased Discrimination	Negatively phrased Discrimination
Arnett8	0.174	2.084	0.629	
Arnett7	0.169	1.778	0.948	
Arnett24	-0.005	0.895	1.748	
Arnett18	-0.206	1.584	2.778	
Arnett25	-0.222	2.315	0.598	
Arnett19	-0.226	2.444	1.153	
Arnett11	-0.231	3.819	0.658	
Arnett14	-0.403	4.727	0.712	
Arnett3	-0.631	3.309	0.660	
Arnett6	-0.688	4.419	0.250	
Arnett16	-0.887	2.298	0.897	
Arnett1	-0.952	3.949	0.429	
Arnett21 (R)	-0.974	2.889		0.520
Arnett20 (R)	-1.017	9.311		8.102
Arnett15 (R)	-1.045	4.528		3.493
Arnett12 (R)	-1.069	2.818		2.574
Arnett13 (R)	-1.082	1.385		0.166
Arnett10 (R)	-1.128	3.547		2.876
Arnett5 (R)	-1.179	2.111		0.502
Arnett26 (R)	-1.187	3.775		2.831
Arnett9 (R)	-1.197	2.467		1.906
Arnett23 (R)	-1.251	1.592		0.482
Arnett22 (R)	-1.263	1.995		1.409
Arnett2 (R)	-1.311	4.309		3.908
Arnett17 (R)	-1.522	3.480		2.767
Arnett4 (R)	-1.893	0.598		1.030

Note.  $n = 750$ . Items are sorted by item difficulty estimates. (R) indicates the item was reverse scored.

**Table 5**

Overall item difficulty and discrimination estimates for centers (4-year sample) based on the bifactor generalized partial credit mode.

Item	Substantive dimension		Method dimensions	
	Difficulty	Discrimination	Positively phrased Discrimination	Negatively phrased Discrimination
Arnett8	-0.171	2.658	0.520	
Arnett7	-0.217	2.162	0.756	
Arnett11	-0.333	4.263	0.463	
Arnett14	-0.428	5.155	0.726	
Arnett25	-0.431	2.080	0.220	
Arnett4 (R)	-0.441	1.688		1.454
Arnett6	-0.498	5.501	0.327	
Arnett3	-0.500	4.617	0.621	
Arnett9 (R)	-0.514	3.763		2.719
Arnett19	-0.521	2.688	1.038	
Arnett1	-0.587	5.533	0.165	
Arnett15 (R)	-0.618	9.037		6.042
Arnett20 (R)	-0.624	9.954		5.945
Arnett22 (R)	-0.635	2.578		1.844
Arnett10 (R)	-0.718	4.464		2.159
Arnett16	-0.760	3.048	1.069	
Arnett12 (R)	-0.775	2.815		1.961
Arnett26 (R)	-0.790	5.047		2.930
Arnett2 (R)	-0.855	4.661		2.238
Arnett5 (R)	-0.973	2.936		0.346
Arnett18	-0.974	2.361	3.004	
Arnett17 (R)	-1.005	2.573		1.565
Arnett21 (R)	-1.044	2.016		0.200
Arnett24	-1.299	0.973	1.503	
Arnett23 (R)	-1.467	1.190		0.139
Arnett13 (R)	-1.470	0.840		0.050

Note.  $n = 1350$ . Items are sorted by item difficulty estimates. (R) indicates the item was reverse scored.

**Table 6**

Associations between the Arnett CIS Measures and caregiver/teacher characteristics

	Homes (2-year sample)	Centers (4-year sample)
Categorical Caregiver/Teacher characteristics means on Arnett CIS Measure and t-test between subgroups		
	<i>M</i>	<i>M</i>
Gender		
Male	.29	.85 <sub>d</sub>
Female	.21	.35 <sub>d</sub>
Race-ethnicity		
Hispanic	-.04 <sub>a</sub>	.48 <sub>e</sub>
Non-Hispanic, Black	-.05 <sub>b</sub>	.18 <sub>efg</sub>
Non-Hispanic, White	.39 <sub>ab</sub>	.38 <sub>f</sub>
Non-Hispanic, Other	.32	.46 <sub>g</sub>
Certification <sup>a</sup>		
No certification	.19 <sub>c</sub>	n/a
Certification	.41 <sub>c</sub>	n/a
Continuous Caregiver/Teacher characteristics and Pearson correlations		
Age	.01	.08 <sup>*</sup>
Experience (years)	.01	.02
Education	.24 <sup>*</sup>	.20 <sup>*</sup>
Certification <sup>b</sup>	n/a	.11 <sup>*</sup>
Group size	-.10 <sup>*</sup>	.06
Child:caregiver ratio	-.07	-.05

Note.  $n = 650$  at 2 years.  $n = 1000$  at 4 years. n/a = not applicable. a–g values with the same subscript letters differ significantly at  $p < .05$ .

<sup>a</sup>Certification at 2 years is a dichotomous indicator of any certificates in early childhood education or related fields (other areas of education, nursing, social work, or psychology).

<sup>b</sup>Certification at 4 years is a continuous variable the sum of five indicators of whether the teacher had various coursework or credentials in early childhood education.

<sup>\*</sup> $p < .05$  (Pearson correlation coefficient differs significantly from zero).



Standardized regression coefficients for child outcomes predicted by process and structural quality in home-based child care and the child's home (2-year sample).

Table 7

Child outcomes	Home-based child care				Family		
	Process		Structure		Process		
	Arnett CIS Bifactor IRT Measure	Caregiver education	Group size	Child: caregiver ratio	Two Bags Task Stimulating/Sensitive	Two Bags Task Harsh/Intrusive	Family SES composite
Cognitive							
BSF-R mental score	.11*	.04	.03	.03	.26*	-.10*	.23*
Socio-emotional							
BSF-R social competence	.15*	.08	-.05	-.02	.13	-.06	.11
BSF-R emotional and behavioral regulation	.15*	.04	-.05	-.04	.17*	-.06	.17*
BSF-R attention and concentration	.14*	.01	-.04	-.06	.11	-.09*	.21*
Child temperament index	-.05	.01	.01	-.01	-.04	-.02	-.16*
Health							
Child excellent health <sup>a</sup>	-.01	.03	-.02	-.01	.04	-.01	-.01
Summary score: absence of illnesses	.00	-.01	-.15*	-.14*	-.13*	.05	-.13*
No injury that required doctor's visit <sup>a</sup>	.01	.01	-.01	.02	-.02	.01	.01
BSF-R motor score	.02	.01	-.01	-.01	.03	.02	.09

Note.  $n = 650$  in all cases except the BSF-R measures, where  $n = 600$ , and the column labeled *Two Bags Task Total Score*, where  $n = 550$ . Values are standardized regression coefficients from models that adjust for the child, family, community and child care covariates listed in Appendix A (included in online supplementary material). Each cell represents a separate regression, with the child outcome listed in the row and the focal predictor listed in the column. Results weighted by the ECLS-B sampling weight.

<sup>a</sup>For dichotomous outcomes, values are changes in predicted probabilities for a one standard deviation increase in the predictor of interest, centered at the mean, with all covariates at their means.

\*  $p < .05$ .

Standardized regression coefficients for child outcomes predicted by process and structural quality in center-based child care and the child's home (4-year sample).

**Table 8**

Child outcomes	Center-based child care				Family		
	Process		Structure		Process		
	Arnett CIS BIFactor IRT Measure	Teacher education	Group size	Child: teacher ratio	Two Bags Task Stimulating/Sensitive	Two Bags Task Harsh/Intrusive	Family SES composite
Cognitive							
Math composite score	-.01	.03	.03	.08*	.06	-.03	.28*
Reading Composite Score	.03	.04	.07	.07	.10*	.00	.29*
Socio-emotional							
Parent report							
Social competence	.04	-.03	.00	.02	.06	.00	.00
Emotional and behavioral regulation	-.02	.01	.09*	-.02	.03	.00	.09
Attention and concentration	-.02	.06	-.01	.04	-.08	-.02	-.11
Teacher report							
Social competence	.00	-.08	.09	.08	.06	-.05	-.07
Emotional and behavioral regulation	.04	.17*	-.04	-.13*	.02	.01	.03
Attention and concentration	.05	.13*	-.02	.00	-.01	.03	-.04
Health							
Child Excellent Health <sup>a</sup>	.02	.07*	-.01	-.04	.01	.04	.06
Summary Score: Absence of Illnesses	.02	.06	-.09*	-.07	.02	.01	.09
No Injury that Required Doctor's Visit <sup>a</sup>	.06*	.04	-.01	-.01	.01	.03	.02

Note.  $n = 1000$  in all cases except the column labeled *Two Bags Task Total Score* where  $n = 900$  to 950. Values are standardized regression coefficients from models that adjust for the child, family, community and child care covariates listed in Appendix A (included in online supplementary material). Each cell represents a separate regression, with the child outcome listed in the row and the focal predictor listed in the column. Results weighted by the ECLS-B sampling weight.

<sup>a</sup>Fordichotomous outcomes, values are changes in predicted probabilities for a one standard deviation increase in the predictor of interest, centered at the mean, with all covariates at their means.

\*  $p < .05$ .