Early Childhood Research Quarterly 24 (2009) 121–132

Contents lists available at ScienceDirect

Early Childhood Research Quarterly

Examining the psychometric properties of the Infant–Toddler Environment Rating Scale-Revised Edition in a high-stakes context

Rossana Bisceglia[a,*], Michal Perlman[a], Diana Schaack[b], Jennifer Jenkins[a]

[a] Department of Human Development and Applied Psychology, University of Toronto, Toronto, Ontario, Canada M5S 1V6
[b] Erikson Institute, Loyola University Chicago, Chicago, IL 60611, USA

ARTICLE INFO

Article history:
Received 6 October 2007
Received in revised form 9 February 2009
Accepted 11 February 2009

Keywords:
ITERS
ITERS-R
Psychometric Properties
Factor Analysis
High-stakes
Staff-to-child ratio
Child care quality
One-classroom score
Multilevel modeling
Intra-class Correlation

ABSTRACT

The psychometric properties of the Infant–Toddler Environment Rating Scale-Revised Edition (ITERS-R) were examined using 153 classrooms from child-care centers where resources were tied to center performance. An exploratory factor analysis revealed that the scale measures one global aspect of quality. To decrease redundancy, subsets of items were selected randomly and by experts who rated items according to ease of administration and importance to quality. The shorter subsets demonstrated good discriminant validity, adequate to good psychometric properties, and high associations to the full ITERS-R score. They also demonstrated similar associations to staff education and staff-to-child ratio, as the full instrument. The best assessment of quality was demonstrated by the shortened subset that included items that assess both structural and process features of quality. Multilevel-analyses indicated that classrooms from the same providers score more similarly on ITERS-R than classrooms from other providers. The implications for using the ITERS-R in high-stakes contexts are discussed.

More and more infants and toddlers are being cared for outside of their homes. In 2005, approximately 20% of American children under the age of two attended regulated early child-care programs (America's Children in Brief: Key National Indicators of Well-Being, 2006). Research on the effects of child-care on children's development has shown that the quality of care matters to children's well being. Higher quality care is associated with positive developmental outcomes, both in the short-term (Burchinal, Roberts, Nabors, & Bryant, 1996; Dunn, 1993; Peisner-Feinberg & Burchinal, 1997) and long-term (Broberg, Wessels, Lamb, & Hwang, 1997; Campbell & Ramey, 1995; Currie & Thomas, 1995; Howes, 1988; Peisner-Feinberg et al., 2001), encompassing a wide range of developmental domains including cognitive functioning, language development, social competence, and emotional adjustment (e.g. Howes, 1988; National Institute of Child Health and Human Development (NICHD) Early Child Care Research Network, 2000; Peisner-Feinberg et al., 2001). Despite the consensus that quality matters, the definition and measurement of quality remain elusive.

In general it is agreed that quality includes both "structural features" such as staff-to-child ratio and group size, and "process features" which involve the interactions and experiences that children have in the child-care center (Goelman et al., 2006). Measurement approaches that attempt to assess overall or global quality include indicators of both structural and process features. The *Infant–Toddler Environment Rating Scale* (ITERS; Harms, Cryer, & Clifford, 1990) is a widely used instrument in research on child care quality (Blau, 2000; Campbell & Milbourne, 2005; Gevers Deynoot-Schaub & Riksen-

* Corresponding author.
E-mail address: rbisceglia@oise.utoronto.ca (R. Bisceglia).

0885-2006/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ecresq.2009.02.001

Walraven, 2006; Goelman et al., 2006; Herrera, Mathiesen, Merino, & Recart, 2005; Scarr, Eisenberg, & Deater-Deckard, 1994; van Ijzendoorn, Tavecchio, Stams, Verhoeven, & Reiling, 1998). The instrument assesses overall or global quality of the classroom environment for children up to 30 months of age. This study presents the psychometric properties of the revised edition of the ITERS (ITERS-R; Harms, Cryer, & Clifford, 2003).

Research on the measurement properties of instruments such as the ITERS-R has become increasingly important given the recent funding initiatives in the United States and to a lesser extent in Canada to motivate improvements in the child-care experience (Zellman & Perlman, 2008; Zellman, Perlman, Le, & Setodji, 2008). These initiatives tie the funding that a child-care center receives to the facility's scores on quality measures. Hence, scores on instruments such as the ITERS-R may increase or decrease a facility's funding. Other initiatives include tying teacher salaries and bonuses to classroom scores on the evaluation instrument, and making scores public so that parents can use them in selecting care for their children.

To our knowledge, no research has examined the validity of using the ITERS-R to make high-stakes decisions. Yet, according to the *Standards for Educational and Psychological Testing* (American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME), 1999) instruments should be administered in contexts and for purposes for which they have been validated. A fundamental purpose of test validation is to examine the positive and negative consequences of test administration, and to determine that specific benefits will likely be realized by the administration of the test.

It is unknown whether the abovementioned practices of tying funding and salaries to ITERS-R scores improve the child-care experience. Conversely it is possible that the contingencies may motivate some child care providers to introduce facile changes to the center, such as modification to the space and furnishing and the arrangement of the classroom, in order to increase their scores on the instrument. If the scale does not measure distinct aspects of quality, such changes may increase the facility's overall score while possibly over-estimating the quality provided by the center. Indeed, research on high-stakes testing in school settings shows that as teachers learn the assessment systems, they modify their classroom curriculum and arrangement in keeping with the content of the evaluation tool (Corbett & Wilson, 1991; Shepard & Dougherty, 1991). Although this may result in improvements in the teaching of certain constructs and skills, concepts that may be more difficult to measure or that are not consistent with the operational definition of quality incorporated in the instrument may not receive adequate attention. We infer from this research that child care providers may respond similarly as elementary and secondary teachers to such high-stakes contexts, rendering the psychometric properties of the instrument particularly important.

Further, the lengthy administration of the ITERS-R, which spans from 4 to 5 h per classroom, may limit the use of the instrument by large-scale agencies, for large research projects and community uses, which require time-efficient instruments. The lengthy administration also raises questions with respect to the allocation of funds and the cost associated with the assessments. Perhaps the funds used in assessment would be better spent on quality improvements in centers that have limited resources. Identifying strategies to reduce evaluation cost can increase the use of the instrument in both research and applied settings.

One strategy to reduce assessment burden could be to administer the instrument to fewer classrooms per center. Research on the psychometric properties of the Early Childhood Environment Rating Scale-Revised (ECERS-R; Harms, Clifford, & Cryer, 1998) which is a sister measure to the original ITERS used in classrooms serving preschool-age children, indicates that classrooms from the same provider tend to score similarly to one another on the instrument compared to classrooms from different providers (see Perlman, Zellman, & Le, 2004). This suggests that the assessment of a few classrooms per center may be as reliable as assessing all classrooms from the same provider. In this paper, we evaluate the validity of this strategy by examining the extent to which the variance in classrooms' ITERS-R scores is attributable to provider membership. To our knowledge no research has empirically examined this question.

An alternate evaluation strategy could be to administer a shorter version of the ITERS-R. However, this strategy would only be valid if the instrument measures one global factor and if the items are highly correlated with one another which would suggest that a shorter version of the scale may capture the construct just as well as the longer version.

In the present study, we report results from an exploratory factor analysis based on the ITERS-R scores from 153 classrooms where funding allocation is tied to the scale's scores. We assess the validity of several cost-saving strategies for assessment, in a similar way as Scarr and colleagues did for the ITERS. Scarr et al. (1994) examined the psychometric properties of three subsets of 12 randomly selected items of the ITERS. The subsets demonstrated psychometric properties similar to the full scale. The authors concluded that the single dominant factor assessed by the full ITERS could be measured with good reliability and validity by using a shorter version. Further, Scarr et al. (1994) suggested that *less* than 12 items may be sufficient in capturing overall quality, and that researchers could select items that are consistent with their own theoretical and methodological purpose. Similar findings were reported for ECERS-R (Perlman et al., 2004). Based on those findings several research efforts have used a shortened version of the ECERS-R (Karoly, Ghosh-Dastidar, Zellman, Perlman, & Fernyhough, 2008). In this paper, we examine the validity of the suggestions made by Scarr et al. (1994) by comparing the psychometric properties of shorter subsets to the full ITERS-R instrument.

We also examined the extent to which the subsets, classified classrooms to the same quality categories as the full instrument. To further test the validity of the subsets, we examined the associations among the score of the full instrument and each subset, and the extent to which the full instrument and the subsets showed similar associations to regulatable indicators predictive of quality. If the subsets demonstrate comparable psychometric properties to the full instrument, researchers and policymakers may choose to administer shorter subsets to reduce time and cost associated with evaluation, while maintaining reliability. Since becoming a reliable rater on the full instrument requires time and frequent

checks to establish and maintain inter-rater agreement, a shorter instrument may facilitate use for research and applied purposes.

1. Method

1.1. Sample

The sample consisted of 153 classrooms from 59 Colorado child-care centers. Staff from Qualistar Early Learning, a non-profit quality improvement agency in Denver, Colorado collected the data from July 2007 to February 2008. Centers were involved in a variety of child-care quality improvement efforts in Colorado and many had been previously assessed using the ITERS. The majority of infant/toddler classrooms in this sample were participating in a public policy initiative in Colorado aimed at improving child care quality and children's school readiness skills. Forty-one (27%) classrooms sampled were located in 14 (24%) programs that were accredited through the National Association for the Education of Young Children (NAEYC). Centers were eligible to participate if they were located within close proximity to underperforming elementary schools and served at least 50% of children who received child care subsidies. Of the classrooms sampled, 30% identified themselves as for-profit centers and 70% identified as not-for-profit; 29% were located in rural regions of Colorado while 71% were located in urban or suburban areas. While this was a baseline ITERS-R administration for all of the classrooms in the sample, 84% of the classrooms had previously been administered the original ITERS. The number of past ITERS observations ranged from 1 to 6 previous administrations.

Classroom assessments were conducted during a four-to-five hour observation period per class. Qualistar Early Learning employed a total of 11 raters who were experienced early childhood education practitioners and who had an associate degree or above. Two raters who worked for an independent agency were trained in North Carolina in programs made available through the authors of the ITERS/ITERS-R. These raters were certified as "Gold Standard" and provided the initial training and drift assessments for all the 11 Qualistar Early Learning data raters in the current study.

The 11 raters collected ITERS-R data and measures of child care structural quality. Raters were assigned to classrooms based on their home proximity to the child care center. Only raters who had not conducted an observation in the center within the past 24 months were eligible to conduct these assessments. To accommodate a request from center directors that the number of raters in each center be minimized, one rater was assigned to collect all ITERS-R data within each center.

Following the reliability procedures developed by the authors of the ITERS-R (see ITERS-R Inter-rater Reliability Sheet (2003) for an example of the scoring sheet) inter-rater reliability was calculated by comparing the raters' score to the "Gold Standard." A disagreement of one point was considered a match (e.g., for reliability purposes, if a rater scored an item as a 3 while the "Gold Standard" scored the same item a 4, their scores were considered a match). Raters who were previously considered trained and reliable on the original ITERS were required to achieve 85% agreement within one point of the "Gold Standard" on each of the items on the ITERS-R. Raters who had not been trained on the ITERS-R were required to achieve 85% agreement within one point of the "Gold Standard" on each ITERS-R item on three consecutive administrations of the ITERS-R. The agency responsible for inter-rater reliability ensured that every rater met the 85% agreement standard, on each item. Unfortunately, it was only noted whether raters achieved this standard. All the raters that were used to collect the data we present met or exceeded the agreement standard. Because the actual rates of item agreements were not available to us, we were unable to compute Cohen's kappa or other statistical measures of inter-rater reliability. Rater agreement was assessed on 18% of the sample and re-examined every 10th observation for each rater to control for rater drift. Drift assessments occurred on average every four months. All raters were reliable with the "Gold Standard" throughout the drift assessments.

These ITERS-R ratings are made public and reimbursement rates and other key "rewards" are attached to the ratings. Given the high stakes associated with the ratings, Qualistar Early Learning is very careful to meet the standards outlined by the measure developers on conducting these assessments reliably. Qualistar ensures that data collectors meet or exceed the reliability standards outlined by the authors of the measure. Using ITERS-R scores collected in a "real-world" context increases our confidence in the ecological validity of our findings.

1.2. Measures

1.2.1. The Infant/Toddler Environment Rating Scale (ITERS-R)

The ITERS-R consists of 39 items organized under seven distinct categories: Space and Furnishings (5 items), Personal Care Routines (6 items), Listening and Talking (3 items), Activities (10 items), Interaction (4 items), Program Structure (4 items), and Parent and Staff (7 items). Data were not collected on the Parent and Staff scale as the items comprising this scale rely heavily on staff-reports rather than observed information. Self-report items were deemed less appropriate given the high-stakes associated with these scores and the known limitations of informant bias (Achenbach, McConaughy, & Howell, 1987). Deletion of the Parent and Staff scale reduced the number of items to 32.

ITERS-R items are rated on a 7-point scale; values 1–3 describe 'inadequate' and 'minimal' care, respectively; values 4–5 describe 'good' and values 6–7 describe 'excellent' quality (Harms et al., 1990). We computed the following scores for inclusion in different analyses as detailed in the subsequent sections of the paper: The 'Full-ITERS-R score' was computed by averaging the scores on all the items for each classroom, a 'provider-level mean score' was computed by averaging the

Table 1
Means, standard deviations and ranges of ITERS-R scores and regulatable measures of quality.

	N	Mean	SD	Minimum	Maximum
Staff-to-child Ratio	153	4.03	1.39	1	8
Post Secondary ECE Credits	153	8.82	11.11	0	51
Staff Education Level	153	1.95	1.64	0	6
Full-ITERS-R score	153	4.9	1.00	2	7
One-Classroom mean score	153	4.93	1.00	2.04	6.85
Provider-level mean score	153	4.93	.915	2.57	6.39

ITERS-R scores across all classroom *within* a center, last a mean score for each subscale was computed by averaging the classrooms' sores on the respective items per scale.

1.2.2. Regulatable quality indices

The following regulatable indices were assessed in this study: classrooms' staff-to-child ratio data, level of education, and number of ECE credits earned. These indices are representative of the regulatable variables evaluated in past research (Scarr et al., 1994) and have shown associations to child-care quality (NICHD, 2000).

Ratio data were obtained through a sign-in/out method. Parents were asked to sign their children into and out of a classroom each morning and evening. Providers were asked to sign-in and out as they arrived and left each classroom during the day. If children or staff left the classroom for non-programmatic reasons, (e.g. to move children to meet better ratios) they signed out of the classroom. Staff and children were not asked to sign in and out if they moved for programming reasons such as going to the playground, to the bathroom, or to another room for instructional purposes. A minute-by-minute adult to child ratio was calculated and averaged across a two-week period to be used in the analysis. For more information on this method see Le, Perlman, Zellman, and Hamilton (2006).

Training and formal education data were collected from the child-care providers for each staff member that worked in the classroom at least 30% of the time. Providers submitted formal training certificates earned over the previous three years and formal transcripts documenting their formal college coursework. Raters assessed transcripts and coded each provider's education level on a scale of 0–8; a point of 0 indicates completion of less than 45 training hours taken in the last three years; 1 indicates completion of 45 or more training hours in the last three years; 2 equals completion of six credits in ECE; 3 indicates completion of a Child Associate Degree (CDA); 4 represents completion of 24 credits in ECE; 5 equals completion of an Associates Degree; 6 refers to completion of a Bachelors Degree; 7 indicates completion of a Masters Degree; and 8 equals completion of a Ph.D. The total number of ECE credits taken for each provider was also calculated. The average education level and number of ECE credits taken for all staff members assigned to a classroom was calculated and used in the analyses.

2. Results

2.1. Descriptive statistics

In Table 1 we present the descriptive statistics of the 'Full-ITERS-R score' which ranged from 2 to 7, with a mean of 4.9. This mean falls within the 'good' category using the classification system developed by Howes, Phillips, and Whitebrook (1992). The high mean is probably explained by the fact that many of the providers in this study had participated in quality improvement initiatives and had been previously assessed for quality. On average, teachers had few ECE credits ($M = 8.82$) and the average education level for staff in each classroom also tended to be low with a mean of 1.95 which approximates completion of 6 ECE credits over the past 3 years. However, the range varied widely, with some classrooms having an average education level of less than 45 training hours while others had an average education level of a Bachelors degree. The staff-to-child ratios in the classroom tended to be small with a mean of 4 children per staff member, but ranged widely from 1 to 8 children per staff member.

ITERS-R items with more than 10 percent missing values were excluded from the analyses; this led to the deletion of the following items 'Group play activities' ($M = 5.25$, $SD = 2.43$, $N = 115$), 'Use of TV/Video/Computers' ($M = 5.18$, $SD = 2.20$, $N = 22$), 'Blocks' ($M = 4.55$, $SD = 1.95$, $N = 133$), 'Art' ($M = 4.4$, $SD = 2.69$, $N = 132$), 'Sand and Water Play'($M = 106$, $SD = 2.21$, $N = 106$). The missing data were a result of the items being rated as 'Not Applicable' as they did not apply to classrooms that cared for very young infants. The 'Provision for special needs' item was deleted as 91% of the centers in our sample did not care for children with special needs.

The distribution of each item was assessed for normality. In general, scores on items dealing with personal care routines concentrated at poor quality levels (e.g. scores of 1–2) while scores on items assessing peer and adult–child interactions, use of language and books, classroom organization, structure, supervision, and availability of furniture, concentrated at high quality levels (e.g. scores of 6–7). Items that measured the provision for relaxation and comfort, the opportunity to learn about nature and science, and to engage in fine motor and music activities demonstrated properties with a bimodal distribution whereby the majority of the scores were either in the poor range (scores of 1–2) or the positive range (scores of 6–7) with infrequent or no endorsements of median values (e.g. 3–4). Items with poorly distributed properties were transformed into

Table 2
Item loadings of the exploratory factor analysis ($N = 153$).

Items	Factor 1	Factor 2	Factor 3
Indoor Space	.45	.09	.12
Furniture for Routine Care and Play	.47	.24	.01
Provisions for Relaxation and Comfort	.54	−.06	−.18
Room Arrangement	.46	.26	−.23
Display for Children	.52	.26	−.07
Health Practices	.40	.35	.65
Safety Practices	.53	.34	.20
Greeting/Departing	.43	−.02	.23
Meals/Snacks	.43	.42	.46
Nap	.42	.30	.11
Diapering/Toileting	.42	.41	.50
Helping Children Understand Language	.76	−.41	.13
Helping Children Use Language	.74	−.34	.14
Using Books	.65	.10	−.14
Fine Motor	.65	−.24	−.14
Active Physical Play	.64	.27	−.23
Music and Movement	.57	−.20	−.17
Dramatic Play	.66	.23	−.36
Nature/Science	.49	.48	−.24
Promoting Acceptance of Diversity	.47	.23	−.51
Supervision of Play and Learning	.65	−.24	.16
Peer Interaction	.66	−.43	.04
Staff-Child Interaction	.72	−.47	.17
Discipline	.67	−.48	.14
Schedule	.73	−.11	−.12
Free Play	.62	.19	−.29

their log. All analyses were conducted separately on the original and transformed items; however, no differences were found therefore we present the results obtained using the original items.

2.2. Psychometric properties of the ITERS-R

We examined the correlations among the individual items and subscales to substantiate the presence of the six distinct a priori subscales described by the authors (Harms et al., 2003). We expected that the subscales would be related to each other given that different dimensions of child-care quality are likely associated. However, if the ITERS-R measures six *distinct* dimensions of child care quality then the magnitude of the correlations should not exceed the modest range. If the magnitude of the correlations is very large, this would suggest that the different scales measure one construct. The average inter-item correlation was .3, and the item-total correlations ranged from .38 to .68, with a median of .50. These are considered moderately high associations given that individual items are not as reliable as composites or subscales (Light, Singer, & Willett, 1990). Subscale correlations were also in the moderate range from .48 to .66 with a median value of .52.

We also calculated the internal consistency estimates (i.e., coefficient alpha) of the a priori subscales and of the full ITERS-R scale; the latter is a measure of the extent to which the items assess the same construct, with low to modest alpha values suggesting the presence of *distinct* categories of quality. The internal consistency estimate of the full ITERS-R was .91 suggesting that the items measure the same construct rather than six distinct dimensions of quality.

2.3. Exploratory factor analysis

To examine the factor structure of the ITERS-R, we subjected the items to an exploratory factor analysis using a principal component extraction method with varimax rotation. The scree plot and the Kaiser criterion (i.e. eigenvalues greater than 1) revealed three factors that accounted for a total of 50% of the variance. The first factor had an eigenvalue of 8.7 and explained 33% of the variance. As shown by Table 2, all items loaded on the first factor with factor loadings ranging from .42 to .76. The second factor had an eigenvalue of 2.4 and explained 9% of the common variance. However, this factor was not readily interpretable as it included positive loadings from items that pertained to personal care routines while negative loadings on 'process features' (use of language, quality of discipline and frequency of peer and adult interaction). Furthermore, all items loaded more strongly on the first and third factor. The third factor had an eigenvalue of 1.9 and explained 7% of the common variance. This factor was identified by 3 of the 6 items from the a priori Personal Care Routines subscale (items: Health Practices, Meals and Snacks and Diapering/Toileting). It is noteworthy, however, that only the 'Health Practices' item clearly loaded more strongly on this factor than on the first factor, while the other two items showed very similar loadings on the first factor.

The higher eigenvalues of the first factor, the lack of interpretability of the second factor and the fact that all items loaded on the first factor, lead us to conclude that a single dominant factor is present.

Table 3
Cronbach's alpha estimates of the full instruments and shortened subsets.

Measure	Cronbach's alpha
Full ITERS-R (26)	.91
Expert-Combined (16)	.90
Quality Subset (9)	.89
Easy-to-Administer Subset (7)	.68
12-item Random Subset 1 (12)	.80
12-item Random Subset 2 (12)	.81
12-item Random Subset 3 (12)	.80
10-item Random Subset 1 (10)	.81
10-item Random Subset 2 (10)	.84
10-item Random Subset 3 (10)	.76
8-item Random Subset 1 (8)	.75
8-item Random Subset 2 (8)	.71
8-item Random Subset 3 (8)	.73

Note. The number of items of each measure is indicated in brackets.

2.4. Psychometric properties of the random subsets, easy-to-administer and quality subsets

To evaluate whether a shorter version of the ITERS-R measures the single dominant factor as well as the full instrument, we evaluated the psychometric properties of shorter subsets. Following Scarr et al.' (1994) procedure, we created 3 subsets of 12-randomly selected items to represent a shorter version of the full ITERS-R. If the 12-item subsets demonstrated comparable psychometric properties in relation to the ITERS-R, we planned to decrease the number of items and create 3 additional subsets of randomly selected items. We continued this procedure until the psychometric properties of the subsets were no longer comparable to the full ITERS-R.

We also assessed two additional subsets of items selected by the authors who served as four child development experts (two Psychology professors who specialize in child development and two senior graduate students in Psychology who have extensive field experience working in early childhood education classrooms as consultants for a variety of quality improvement efforts). One expert-selected subset consisted of 7 easy-to-administer items (i.e., items that are easy to observe such as having a comfortable resting place for children); the second subset included 9 items that were judged by the experts as particularly important for child-care quality. Easy-to-administer items related to the physical environment such as room arrangement, furnishings and pictures displayed. Items identified as important for quality tended to be more process-oriented, focusing on peer, and child-adult interaction and discipline (see Appendix A for a complete list of items).

In Table 3 we present the Cronbach's alpha of the shorter subsets. The three random subsets of 12 items demonstrated good internal consistency estimates of .80–.81; these estimates approximate the Cronbach's alpha of the entire scale of .91. Alpha estimates for the easy-to-administer and quality subsets were .68 and .89, respectively. It is worth noting that 5 of the 7 items comprising the easy-subset derived from the Space and Furnishings a priori subscale which demonstrates the lowest internal consistency than the other scales; the authors of the ITERS-R also reported a Cronbach's alpha of .47 for this a priori subscale (Harms et al., 2003). We also calculated the internal consistency estimates of the quality- and easy-items combined (Expert-combined subset); this resulted in $\alpha = .90$. Therefore, with the exception of the easy-to-administer subset, the shorter subsets demonstrated good internal consistency estimates, with the expert-combined and quality subsets demonstrating almost identical internal consistency estimates as the full instrument. Further, the 12-item random subsets, the Quality and Expert-combined subsets demonstrated similar or higher internal consistency estimates than the a priori subscales: Space and Furnishings ($\alpha = .61$, $N = 5$), Personal Care Routines ($\alpha = .72$, $N = 6$), Listening and Talking ($\alpha = .75$, $N = 3$), Activities ($\alpha = .77$, $N = 6$), Interaction ($\alpha = .84$, $N = 4$), Program Structure ($\alpha = .70$, $N = 2$). Correlations between the five subsets, the combined expert-selected items and the full 'Full-ITERS-R score' proved to be substantial, ranging from .93 to .96. These correlations remained strong after removing common items from the full ITERS-R; the correlations ranged from .76 to .93, with a median of .91. The high correlations suggest that the subsets were psychometrically similar to the full ITERS-R.

The 10-item subsets demonstrated internal consistency estimates of .76–.84, while the 8-item version demonstrated alpha estimates of .71–.75. These subsets also demonstrated high correlations to the 'Full-ITERS-R score' which ranged from .93 to .94 for the 10-items subsets, and .91 to .93 for the 8-items subsets. The magnitude of association remained high after removal of common items from the 'Full-ITERS-R score'; these correlations ranged from .85 to .92 for the 10-items subsets and .82 to .91 for the 8-items subset. Hence, even subsets with as few items as 8 remained strongly correlated to the 'Full-ITERS-R score' and demonstrated adequate internal consistency. The lower alpha of the 8-item subsets may be attributable to a decrease in the number of items which is an important element for adequate alpha estimates (Streiner & Norman, 1989). The strong correlations between these subsets and the 'Full-ITERS-R score' suggest that the subsets captured the same construct as the full ITERS-R.

Table 4
Associations between the subsets, full ITERS-R score and regulatable indicators.

	Full ITERS-R score	12-items Random Subset			Quality	Easy	Expert-combined
		1	2	3			
Full ITERS-R score	–	.93**	.95**	.96**	.90**	.86**	.96**
Staff Education Level	.28*	.29**	.24*	.33 **	.23*	.26*	.26**
Post Secondary ECE Credits	.18*	.19*	.17*	.20*	.13	.18*	.16*
Staff-to-child Ratio	−.33**	−31**	−.36**	−.32*	−.35**	−.29**	−.36**

* Correlations were significant at the $p < .05$ level.
** Correlations were significant at the $p < .001$ level.

2.5. Relationship to other quality indicators

We examined the extent to which the correlations between the 'Full-ITERS-R score' and the following regulatable indicators: staff education level, number of ECE credits and staff–child-ratio were approximated by the subsets. Similar patterns in associations would indicate that the subsets measure the same construct as the full ITERS-R. Associations were examined only for those subsets that were found to have comparable psychometric properties to the full ITERS-R.

As shown in Table 4, significant associations were found between the 'Full-ITERS-R score' and all of the regulatable indices; this pattern of correlations was also approximated by the 12-item random subsets, the quality, easy-to-administer and the expert-combined subset. For example, the magnitude of the correlation between the 'Full-ITERS-R score' and staff education level was .28; similarly, the associations between the subsets and Staff Education level ranged from .23 to .33 in magnitude. To evaluate whether the magnitude of association between the subsets and the regulatable indicators (e.g. association between subset 1 and staff education level) were equivalent to those shown by the full instrument (e.g. association between full ITERS-R score and staff education level), we applied Fisher's *r-to-z* transformation (Fisher, 1915) where the value of the correlation coefficient is transformed into a *z* value. A *z* value of 1.96 would be significant at .05 level; no *z* value reached significance (*z* values ranged from 0 to .89) indicating that the subsets showed a similar relationship to the regulatable indicators as that shown by the full instrument.

2.6. Variance partitioning

We investigated the degree of similarity in the scores of classrooms from the same center using multilevel modeling. An unconditional, two-level model (also called a one-way random effects model) was run as there were two sources of variation (classrooms nested within centers) using the program Mlwin version 2.0 (Rasbash, Steele, Browne, & Prosser, 2004). MLM is a regression-based analytic technique that is ideal for analyzing nested data structures. Variance on the response variable (i.e. ITERS-R scores) is partitioned into within and between provider variance, showing the extent of clustering of classes within centers. Results of the model showed that the within and between provider variances were both significant. Most of the variance (73%) was at the provider level ($\sigma^2 u_0 = .75$, $SE = .16$; $\chi^2 = 21.30$, $p < .001$). This means that classrooms within centers show a high degree of similarity to one another. Only 27% of the variance was at the within provider level ($\sigma^2 e_0 = .27$, $SE = .04$; $\chi^2 = 47.09$, $p < .001$). This resulted in an Intra-class correlation (calculated as the provider level variance divided by the total variance, with a maximum score of 1.0) of .73.

2.7. Discriminant analyses

We performed discriminant analyses to assess the extent to which the shorter subsets would assign providers to the same category as the 'Full-ITERS-R score.' We first categorized the 'Full-ITERS-R score' to create three categories of quality of care, via dummy coding based on the ITERS-R 7-point scale. Classrooms were assigned to the 'poor quality' category if they received a 'Full-ITERS-R score' of less than and equal to 3; those who received a mean score above 3 but below 5 were assigned to the 'average quality' category; classrooms with a mean score between 5 and 7 were considered of 'good quality.'

A requirement of the discriminant analyses is to assign a sample size to the different groups. One option is to assign the groups' prior probabilities of the sample distribution based on the 'Full-ITERS-R score'; this function assumes that the relative sample sizes for the groups in the sample are estimates of population proportions. We employed a second option, which involves assigning equal distributions for each of the groups; this assumes that the distribution of classrooms should be equal across groups. Setting the frequency to equal better reflected the psychometric goals of the study: To evaluate the accuracy at which the subsets assign the classrooms to the same categories as the score based on all the ITERS-R items.

In Table 5 we present the results of the discriminant analyses. Correctly assigned cases appear underlined on the diagonal lines of the table, per random subset. For example, the 12-items Random Subset 1 correctly assigned 8 of the 8 cases of the 'poor' quality group (100%), 40 of 59 of the 'average' (68%) and 80 of 86 of the 'good' quality category (93%). Overall the 12-items random subset 1 correctly assigned 84% of the classrooms to the same quality category as the full instrument. The overall percent of correct assignment is the total number of correct cases (diagonal values, for 12-items random subset 1 = 128) divided by the total number of classrooms in the sample ($N = 153$). Cases that were incorrectly assigned by the

Table 5
Discriminant analyses assessing the accuracy of the 12-items random subsets, expert-selected items, the one-classroom and provider-level score in categorizing classrooms into the same quality categories as the full-ITERS-R score.

Subset	Predictor measure	Group Membership			% correct per subset
		Poor (8)	Average (59)	Good (86)	
12-items Random Subset 1	Poor	8 (100%)	0	0	84%
	Average	11 (19%)	40 (68%)	8 (13%)	
	Good	0	6 (7%)	80 (93%)	
12-items Random Subset 2	Poor	8 (100%)	0	0	91%
	Average	5 (8.5%)	49 (83%)	5 (8.5%)	
	Good	0	4 (5%)	82 (95%)	
12-items Random Subset 1	Poor	8 (100%)	0	0	86%
	Average	7 (12%)	43 (73%)	9 (15%)	
	Good	0	6 (7%)	80 (93%)	
Easy-to-Administer	Poor	8 (100%)	0	0	72%
	Average	13 (22%)	29 (49%)	17 (29%)	
	Good	0	13 (15%)	73 (85%)	
Quality	Poor	7 (88%)	1 (12%)	0	82%
	Average	5 (9%)	38 (64%)	16 (27%)	
	Good	0	6 (7%)	80 (93%)	
Expert-combined	Poor	8 (100%)	0	0	88%
	Average	5 (9%)	43 (73%)	11 (18%)	
	Good	0	3 (3%)	83 (97%)	
One-classroom mean score[a]	Poor	6 (75%)	2 (25%)	0	80%
	Average	9 (15%)	43 (73%)	7 (12%)	
	Good	0	13 (15%)	73 (85%)	
Provider-level mean score[b]	Poor	7 (88%)	1 (12%)	0	78%
	Average	13 (22%)	33 (56%)	13 (22%)	
	Good	0	7 (8%)	79 (92%)	

Note. Correctly assigned cases appear underlined on the diagonal lines of the table.

[a] 'One-classroom mean score' is the ITERS-R score of one classroom randomly selected from a provider.

[b] 'Provider-level mean score' is the average ITERS-R score of all classrooms from the same provider.

subsets appear without the underline. For example, for the 12-items random subset 1, 0 of the 'poor' quality classrooms were incorrectly assigned to either the 'average' or 'good' category. While 11 of the 59 'average' quality classrooms were incorrectly assigned to the 'poor' quality (19%) and 8 to the 'good' quality category (13%), and 6 of the 86 'good' quality classrooms were incorrectly assigned to the 'average' quality group (7%). Overall, the subsets assigned 72–91% of classrooms to the same category as the 'Full-ITERS-R score.'

We also combined the easy-to-administer and quality subsets (Expert-combined) to assess the discriminant validity of this combination of items. On average, the Expert-combined subset assigned 88% of the classrooms to the same category as the full instrument. Therefore, with the exception of the easy-to-administer subset, on average, the subsets demonstrated acceptable discriminant validity and assigned classrooms to the same quality category as the full instrument.

We further examined whether using the score of *one* classroom randomly selected per center would suffice to predict the quality category of all classrooms from the same center. We first randomly selected one classroom per provider. Second, we derived a 'One-classroom mean score' by computing the average ITERS-R score of a randomly selected classroom from each provider. Third, we assigned the 'One-classroom mean score' to *each* classroom nested within the respective center. Classrooms were assigned to quality categories as aforementioned. As shown in Table 5, in relation to the 'Full-ITERS-R score' overall the 'One-classroom mean score' correctly assigned 80% of classrooms and correctly assigned 6 of the 8 classrooms of the 'poor quality' group (75%), 43 of 59 of the 'average quality' group (73%) and 73 of 86 of the 'good quality' category (85%).

Next we examined the discriminant validity of the 'provider-level mean score' (the average ITERS-R score across *all* classrooms *within* a center) using the aforementioned procedures. This analysis was performed to further examine whether the 'One-classroom mean score' provides similar information as the assessment of all classrooms from the same center.

In relation to the full ITERS-R score, overall the 'provider-level mean score' accurately assigned 78% of classrooms, and correctly assigned 88% of 'poor' quality classrooms, 56% of the 'average' quality and 92% of the 'good' quality classrooms. Hence, on average the 'One-classroom mean score' correctly assigned classrooms to the same quality categories as the 'Full-ITERS-R score' and the 'provider-level mean score.' These results suggest that the assessment of one classroom per center may enable researchers to make reasonably accurate inferences about quality in the infant/toddler rooms in a given center. However, since the classification was not perfect, the assessment of one classroom per centre is not appropriate for evaluation purposes, especially in high-stakes settings.

3. Discussion

This study evaluated the psychometric properties of the ITERS-R; a widely used scale for the measurement of child-care quality. Results from an exploratory factor analysis indicate that all items load on a single factor, suggesting that the instrument does not measure six distinct dimensions of quality (recall that data on the 7th scale 'Parents and Staff' were not available). The high internal consistency estimates of the full instrument and strong inter-item correlations further suggest that the scale measures a single construct.

To the best of our knowledge, only one other study to-date has examined the factor structure of the ITERS-R (see Hestens, Cassidy, Hegde, & Lower, 2007). The authors assessed a 7-, 4- and 3-factor model; although all models showed inadequate to modest fit indices, the four-factor model was retained as it demonstrated the best fit in relation to the other models and the following easily interpretable factors: (1) Materials/Activities, (2) Safety/Organization, (3) Language/Interactions, and (4) Parents/Staff. Given that our findings are discrepant with those of Hestens et al. (2007) and that Hestens and colleagues report modest fit indices, it will be important in future work to continue to examine the factor structure of the ITERS-R and to attempt to identify sampling differences (e.g. high-stakes settings) across studies that may result in the emergence of different factor structures.

Consistent with Hestens et al. (2007) findings, it is possible that a 'Parent and Staff' factor may have emerged if our exploratory factor analysis had included items from the 'Parent and Staff' scale. However, it is noteworthy that measures of parent involvement in child care settings tend to exhibit limited variability and generally positive scores (e.g. see Zellman & Perlman, 2006). In fact, the developers of Qualistar Early Learning's rating system decided not to administer the 'Parent and Staff' items precisely because these yielded little variability in earlier analyses. We therefore speculate that the 'Parent and Staff' scale on the ITERS-R may inflate scores across providers without changing their rank ordering by quality. Also, it is unlikely that their inclusion would have changed the interrelationships among the items that were examined in this study.

In addition to the exclusion of the 'Parent and Staff' items, six additional items were dropped from the analyses as these items did not apply to the sampled classrooms. For example, the 'Provision for Special Needs' item did not apply to 91% of the sample as the classrooms did not care for children with special needs. The 'Use of TV/Video/Computers' item did not apply to 86% of the sampled classrooms as the ITERS-R scoring instructions indicate that this item should be scored as not applicable if classrooms do not use such products. Last, the ITERS-R scoring instructions indicate that the items 'Blocks', 'Group Play Activities', 'Art', and 'Sand and Water Play' do not apply to classrooms that care for children younger than a specified age. For example, the 'Sand and Water Play' item is only applicable to classrooms that serve children who are older than 18 months of age. Hence, the availability of data on the dropped items seemed to be a function of the criteria specified in the ITERS-R scoring instruction (e.g. age group of the children) rather than differences in quality across classrooms. We therefore speculate that the inclusion of these items would not have changed the factor structure reported herein.

An additional limitation concerns the modest sample size of this study. This did not permit us to perform both exploratory and confirmatory factor analyses, which would have allowed us to identify and test multi-factorial solutions. Despite these limitations, our findings are consistent with the ways in which agencies in North America use the ITERS-R to calculate the quality rating. Across all states, the quality rating is based on the ITERS-R total score; hence, the instrument is used as if it has one general dimension of quality.

As our factor analysis indicated the presence of a single dominant factor, we examined whether a shortened version of the instrument would assess the single factor just as well as the full instrument. The use of shorter subsets would relieve the financial and time-burden of the lengthy assessment required for the full instrument thereby potentially increasing its use for both research and applied purposes.

Two subsets, the 'Quality' and 'Expert-Combined' subsets, demonstrated almost identical psychometric properties as the full instrument. The 'Quality' subset consisted of items that relate to 'process features' of child-care quality such as opportunities for peer-peer and adult–child interaction. The 'Expert-Combined' subset consisted of items from both the 'Quality' and 'Easy-to-Administer' subset with the latter measuring 'structural features' of quality such as the availability of classroom materials, the quantity and organization of furniture, etc. Random subsets of at least 12 items also demonstrated acceptable internal consistency estimates; this suggests that the psychometric properties will not be markedly compromised by utilizing a smaller number of items (e.g. 12 items).

Furthermore, the results from the discriminant analyses indicate that the shorter subsets (excluding the 'Easy-to-Administer' version) demonstrated high predictive validity in assigning 82–91% of classrooms to the same quality categories as the full instrument. It is noteworthy, however, that the subsets' classification was not perfect. This indicates the loss of some predictive validity in assigning classrooms to the same quality category as the full instrument.

The lowest predictive validity was demonstrated by the 'Easy-to-Administer' subset, which may be related to the lower internal consistency of the scale. These results indicate that the assessment of classrooms by the administration of items chosen only because they are easy-to-observe (e.g. room arrangement, furnishings and displays) is likely to produce a *less* reliable estimate of the classroom quality, than 12-items randomly selected and the quality items that are process-oriented, focusing on peer, and child-adult interaction and discipline. It is noteworthy that these findings are *inconsistent* with those shown by a study that examined the psychometric properties of the ECERS-R (Perlman et al., 2004), a sister measure to the ITERS-R. Perlman and colleagues found that the sole use of the easy-to-administer items classified, on average, 94% of classrooms to the same quality category as the full instrument. This suggested that the use of only easy-to-administer items provided similar information to that obtained from the full ECERS-R while reducing time and cost burden. Others, however,

because of the identification of two factors on the ECERS-R (Cassidy, Hestenes, Hegde, Hestenes, & Mims, 2005) suggest that a shorter subset consisting of a *combination* of items that load on the two factors, may be ideal for simulating the full ECERS-R.

Although findings from the present study suggest that the ITERS-R measures a single dominant factor, similar to Cassidy et al. (2005) we also found that the *combination* of easy-to-administer and process-oriented items (Expert-combined subset) demonstrated almost identical psychometric properties as the full scale, and good predictive validity in assigning 88% of classrooms to the same quality category as the full instrument. Furthermore, the 'Expert-combined' subset demonstrated almost perfect associations to the 'Full-ITERS-R score' and a similar magnitude of association to the regulatable indicators as that seen for the full instrument. Altogether, these findings indicate that the 'Expert-combined' subset is likely to measure the underlying construct as well as the full instrument. These findings also suggest that while it is possible to reduce the number of items, the best assessment of child-care quality is made by the shortened versions that contain items that assess *both* process and structural features of quality.

The 'Expert-combined' subset, may be especially useful for research purposes as one will need to establish reliability on a fewer number of items relative to the full instrument, however, it may not suffice for quality improvement efforts which require longer and exhaustive inventories in order to provide caregivers with detailed feedback on a wider range of domains that require improvement. The reduction in the number of items will reduce the burden on observers and should make it feasible to reduce the duration of the observation period. However, given the inclusion of process-oriented items this version of the ITERS-R will still require a non-trivial observation window. Also, until the associations between the 'Expert-combined' subset and child outcomes are examined it is not known whether the subset sufficiently measures those aspects of the classroom environment that are important for optimal child development.

In addition to examining the validity of using shorter subsets of the full instrument, we also examined the validity of administering the full ITERS-R to only one-classroom in a center instead of assessing every classroom. Our results from the multilevel analyses indicated that 73% of the variance in ITERS-R scores in our sample was at the provider level; thus, classrooms from the same child-care center scored very similarly to one another when compared with classrooms from other centers. Moreover, the results of the discriminant analyses showed that the score of one classroom randomly selected from each provider, on average, classified classrooms in the same category as the 'Full-ITERS-R score' and the providers' overall mean score. Therefore, it appears that the assessment of one classroom per center may suffice in estimating the overall quality provided by all the classrooms within a center. Researchers may choose to randomly select a few classrooms within centers instead of evaluating each classroom. In practice, the assessment of a few classrooms within a center may suffice in informing general improvement efforts.

Two caveats of the multilevel analyses are important to note. First, since one rater assessed every classroom from the same provider it is possible that the rater scored classrooms from the same center similarly due to a 'halo effect' or other bias. Future research on the degree of similarity among classrooms from the same provider should employ multiple raters so that classrooms within centers are assessed by different raters. It is noteworthy that all raters employed in this study met the reliability criteria outlined by the developers of the ITERS-R which was tested at regular intervals throughout the data collection period. Such procedures are implemented specifically to counter the potential source of rater-bias. Given that the rater agreement and the degree of within-provider similarity were high (.73) we think that it is unlikely that this finding is completely attributable to rater-bias.

Second, it should be noted that the degree of similarity between classrooms from the same center was not perfect. Therefore, it is unlikely that it will be appropriate to base high-stakes decisions, such as increasing or decreasing teacher's salaries, on the assessment of only a few classrooms per center. Such reduced measurement may only be advisable when the stakes for individuals are low.

3.1. Policy implications

The finding that the ITERS-R measures a dominant factor may indicate that different aspects of child-care quality cluster together so that providers who score high on one dimension are also likely to be strong in other areas. These findings are not unique to the ITERS-R. Other widely used quality instruments such as the ECERS (Scarr et al., 1994) and ECERS-R (Perlman et al., 2004; but see Cassidy et al., 2005) have demonstrated similar results as the ones reported in this manuscript.

The finding that the above-stated instruments measure a single factor may also imply *construct underrepresentation* (AERA et al., 1999) of the scales which refers to the degree to which an instrument fails to capture important features of the phenomena under investigation. Examination of the ITERS-R items suggests that the majority of the scale's content pertains to the physical aspects of the classroom, such as the quality of the furniture, the organization of the classroom and program and the quantity of various materials. Fewer items capture process variables such as the quality of child-to-child and child-to-adult interactions.

The redundancy in the scale and the possible *construct underrepresentation* is particularly concerning if the instrument is used in contexts where the stakes are high, such as in settings where salaries and center's funding are contingent on the instrument's overall score. Research on high stakes testing suggests that these contingencies may motivate providers to introduce changes that target specific content of the scale, in order to increase the overall score. Since the ITERS-R items are highly correlated with one another, changes to a few items may increase the overall score thereby erroneously estimating the quality provided by the center. Furthermore the lower internal consistency estimates of the easy-to-observe items indicates

that these items are less reliable measures of the child center quality, yet they are likely to be the ones selected for modification as they can be changed more easily than items that capture the quality of interactions between staff and children.

Data for this paper came from Qualistar Early Learning, an organization that collects the ITERS-R for monitoring and quality improvement purposes. Use of the ITERS-R in such high stakes settings is growing. Analysis of such data is valuable because it provides information about the measure that is ecologically valid and therefore useful in applied as well as research contexts. Consistent with the best practice recommendations specified in the *Standards for Educational and Psychological Testing* (AERA et al., 1999) for testing in high-stakes settings, our findings suggest that in settings where the stakes are high, it is best to use the full ITERS-R in conjunction with other quality-measures. This will ensure that the multidimensional aspect of quality is more accurately reflected in the providers' and classrooms' ratings. Regulatable indicators such as staff-to-child ratios and staff members' credentials are the most widely used quality indices in conjunction with evaluation instruments. Yet, the validity of regulatable variables is questionable as these are often collected via self-report (Scarr et al., 1994). Additional instruments of child-care quality that capture the multidimensionality of quality are greatly needed, especially given their increasing use in high-stakes settings.

Acknowledgements

The authors are grateful to the staff from Qualistar Early Learning center in Denver, Colorado for allowing us to use their child care center data.

Appendix A. List of ITERS-R items per subset and entire scale

Complete Item pool	Subset (12)			Subset (10)			Subset (8)			Easy	Quality	Expert-combined
	1	2	3	1	2	3	1	2	3			
Indoor space	√	√	√		√	√	√	√	√	√		√
Furniture for routine care and play	√	√	√						√	√		√
Provision for relaxation and comfort		√		√						√		√
Room arrangement	√							√		√		√
Display for children		√			√					√		√
Greeting/Departing		√	√			√						
Meals/Snacks	√			√		√		√				
Nap	√		√			√			√			
Diapering/Toileting	√					√	√					
Health practices	√	√	√	√								
Safety practices	√	√			√							
Helping children understand language	√			√	√		√				√	
Helping children use language											√	
Using books	√				√		√		√		√	
Fine motor		√	√	√		√		√		√		
Active physical play			√		√							
Music and movement		√		√	√		√	√	√			
Dramatic play				√	√							
Nature/Science			√	√			√	√				
Promoting acceptance of diversity	√		√		√		√		√	√		
Supervision of play and learning	√	√				√		√			√	√
Peer interaction			√	√		√	√		√		√	√
Staff-child interaction			√	√		√		√			√	√
Discipline		√							√		√	√
Schedule			√		√						√	√
Free play		√				√					√	√

References

Achenbach, T. M., McConaughy, S. H., & Howell, C. T. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin*, *101*, 213–232.

America's Children in Brief: Key National Indicators of Well-Being. (2006). Retrieved September 17, 2007, from http://www.childstats.gov/americaschildren/.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for Educational and Psychological Testing.* Washington, DC: American Educational Research Association.

Blau, D. M. (2000). The production of quality in child care centers: Another look. *Applied Developmental Science*, *4*, 136–148.

Broberg, A., Wessels, H., Lamb, M., & Hwang, C. (1997). Effects of day care on the development of cognitive abilities in 8-year-olds: A longitudinal study. *Developmental Psychology*, *33*, 62–69.

Burchinal, M., Roberts, J., Nabors, L., & Bryant, D. (1996). Quality of center child care and infant cognitive and language development. *Child Development*, *67*, 606–620.

Campbell, F., & Ramey, C. (1995). Cognitive and school outcomes for high risk African American students at middle adolescence: Positive effects of early intervention. *American Educational Research Journal*, *32*, 743–772.

Campbell, P. H., & Milbourne, S. A. (2005). Improving the quality of infant–toddler care through professional development. *Topics in Early Childhood Special Education (Austin)*, *25*(1), 3–14.

Cassidy, D. J., Hestenes, L. L., Hegde, A., Hestenes, S., & Mims, S. (2005). Measurement of quality in preschool child care classrooms: An exploratory and confirmatory factor analysis of the Early Childhood Environment Rating Scale-Revised. *Early Childhood Research Quarterly*, *20*, 345–360.

Corbett, H. D., & Wilson, B. L. (1991). *Testing, reform, and rebellion*. Norwood, NJ: Ablex.
Currie, J., & Thomas, D. (1995). Does head start make a difference? *The American Economic Review*, *85*, 341–364.
Dunn, L. (1993). Proximal and distal features of day care quality and children's development. *Early Childhood Research Quarterly*, *8*, 167–192.
Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*, 507–521.
Gevers Deynoot-Schaub, M. J., & Riksen-Walraven, J. M. (2006). Peer contacts of 15-month-olds in childcare: Links with child temperament, parent-child interaction and quality of childcare. *Social Development*, *15*(4), 709–729.
Goelman, H., Forer, B., Kershaw, P., Doherty, G., Lero, D., & LaGrange, A. (2006). Towards a predictive model of quality in Canadian child care centers. *Early Childhood Research Quarterly*, *21*(3), 280–295.
Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early Childhood Environment Rating Scale Revised Edition*. New York: Teachers College Press.
Harms, T., Cryer, D., & Clifford, R. (1990). *The Infant and Toddler Environment Rating Scale*. New York: Teachers College Press.
Harms, T., Cryer, D., & Clifford, R. M. (2003). *The Infant/Toddler Environment Rating Scale-Revised Edition*. New York: Teachers College Press.
Herrera, M. O., Mathiesen, M. E., Merino, J. M., & Recart, I. (2005). Learning contexts for young children in Chile: Process quality assessment in preschool centres. *International Journal of Early Years Education*, *13*(1), 13–27.
Hestens, L. L., Cassidy, D. J., Hegde, A. V., & Lower, J. K. (2007). Quality in inclusive and noninclusive infant and toddler classrooms. *Journal of Research in Childhood Education*, *22*(1), 69–84.
Howes, C. (1988). Relations between early child care and schooling. *Developmental Psychology*, *24*, 53–57.
Howes, C., Phillips, D., & Whitebrook, M. (1992). Thresholds of quality: Implications for the social development of children in center-based care. *Child Development*, *63*, 449–460.
Karoly, L. A., Ghosh-Dastidar, B., Zellman, G. L., Perlman, M., & Fernyhough, L. (2008). *Prepared to learn: The nature and quality of early care and education for preschool-age children in California*. Santa Monica, CA: RAND Corporation.
ITERS-R Interrater Reliability Sheet. (2003). Retrieved May 16, 2008 from University of North Carolina at Chapel Hill, The Frank Porter Graham Child Development Institute. http://www.fpg.unc.edu/.
Le, V. N., Perlman, M., Zellman, G. L., & Hamilton, L. S. (2006). Measuring child-staff ratios in child care centers: Balancing effort and representativeness. *Early Childhood Research Quarterly*, *21*(3), 267–279.
Light, R. J., Singer, J. D., & Willett, J. B. (1990). *By design: Planning research on higher education*. Cambridge, MA: Harvard University Press.
NICHD Early Child Care Research Network. (2000). The relation of child care to cognitive and language development. *Child Development*, *71*, 960–980.
Perlman, M., Zellman, G. L., & Le, V. (2004). Examining the psychometric properties of the Early Childhood Rating Scale-Revised (ECERS-R). *Early Childhood Research Quarterly*, *19*(3), 398–412.
Peisner-Feinberg, E. S., & Burchinal, M. R. (1997). Relations between preschool children's child-care experiences and concurrent development: The cost, quality, and outcomes study. *Merrill-Palmer Quarterly*, *43*, 451–477.
Peisner-Feinberg, E., Burchinal, M., Clifford, R., Culkin, M., Howes, C., Kagan, S., et al. (2001). The relation of preschool child-care quality to children's cognitive and social development trajectories through second grade. *Child Development*, *72*, 1534–1553.
Rasbash, J., Steele, F., Browne, W., & Prosser, B. (2004). *A user's guide to MlwiN, Version 2.0*. London: University of Bristol, Centre for Multilevel Modeling.
Scarr, S., Eisenberg, M., & Deater-Deckard, K. (1994). Measurement of quality in child care centers. *Early Childhood Research Quarterly*, *9*, 131–151.
Shepard, L. A., & Dougherty, K. C. (1991, April). *Effects of high-stakes testing on instruction*. Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago, IL.
Streiner, D. L., & Norman, G. R. (1989). *Health measurement scales: A practical guide to their development and use*. New York: Oxford University Press.
van Ijzendoorn, M. H., Tavecchio, L. W. C., Stams, G. J. J. J. M., Verhoeven, M. J. E., & Reiling, E. J. (1998). Quality of center day care and attunement between parents and caregivers: Center day care in cross-national perspective. *Journal of Genetic Psychology*, *159*(4), 437–454.
Zellman, G. L., & Perlman, M. (2006). Parent involvement in child care settings: Conceptual and measurement issues. *Early Child Development and Care*, *176*(5), 521–538.
Zellman, G. L., & Perlman, M. (2008). *Child care quality rating improvement systems in five pioneer states: Implementation issues and lessons learned*. Santa Monica, CA: RAND Corporation.
Zellman, G. L., Perlman, M., Le, V., & Setodji, C. M. (2008). *Assessing the validity of the Qualistar early learning quality rating and improvement system as a tool for improving child-care quality*. Santa Monica, CA: RAND Corporation.