Reliability and Validity of the Infant/Toddler Environment Rating Scale

The ITERS-R is a revision of the widely used and documented ITERS, that is one in a family of instruments designed to assess the overall quality of early childhood programs. Together, with the original instrument, the Early Childhood Environment Rating Scale (ECERS), and the more recent revision of that scale, the ECERS-R, these scales have been used in major research projects in the United States as well as in a number of other countries. This extensive research has documented both the ability of the scales to be used reliably and the validity of the scales in terms of their relation to other measures of quality and their tie to child development outcomes for children in classrooms with varying environmental ratings.

In particular, both the ECERS and ITERS scores are predicted by structural measures of quality such as child-staff ratios, group size, and staff education levels (Cryer, Tietze, Burchinal, Leal, & Palacios, 1999; Phillipsen, Burchinal, Howes, & Cryer, 1998). The scores are also related to other characteristics normally expected to be related to quality such as teacher salaries and total program costs (Cryer et al., 1999; Marshall, Creps, Burstein, Glantz, Robeson, & Barnett, 2001; Phillipsen et al., 1998; Whitebook, Howes, & Phillips, 1989). In turn, rating scale scores have been shown to predict children's development (Burchinal, Roberts, Nabors, & Bryant, 1996; Peisner-Feinberg et al., 1999).

Since the concurrent and predictive validity of the original ITERS is well established and the current revision maintains the basic properties of the original instrument, the studies of the ITERS-R have focused on the degree to which the revised version maintains the ability of trained observers to use the scale reliably. Additional studies will be needed to document the continued relationship with other measures of quality as well as to document its ability to predict child outcomes. A two-phase study was completed in 2001 and 2002 to establish reliability in use of the scale.

The first phase was a pilot phase. In this phase a total of 10 trained observers in groups of two or three used the first version of the revised scale in 12 observations in nine centers with infant and/or toddler groups. After these observations, modifications were made in the revised scale to adjust for issues that arose in the pilot observations.

The final phase of the field test involved a more formal study of reliability. In this phase, six trained observers conducted 45 paired observations. Each observation lasted approximately three hours, followed by a 20-30 minute teacher interview. The groups observed were selected to be representative of the range of quality in programs in North Carolina. North Carolina has a rated license system that awards points for various features related to quality. Centers are given a license with one to five stars depending on the total number of points earned. A center receiving a one-star license meets only the very basic requirements in the licensing law while a five-star center meets much higher standards. For our sample we selected 15 groups in centers with one or two stars, 15 with three stars, and 15 with four or five stars. The programs were also chosen

to represent various age ranges of children served. Of the 45 groups observed, 15 were from groups with children under 12 months of age, 15 from groups with children 12-24 months old, and 15 with children 18-30 months old. The groups were in 34 different centers and seven of them included children with identified disabilities. All centers were in the central portion of North Carolina.

The field test resulted in 90 observations with two paired observations each in 45 group settings. Several measures of reliability have been calculated.

Indicator Reliability. Across all 39 items in the revised ITERS, there are a total of 467 indicators. There was agreement on 91.65% of all indicator scores given by the raters. Some researchers will omit the Parents and Staff Subscale in their work. Thus, we have calculated the indicator reliability for the child specific items in the first six subscales, Items 1-32. The observer agreement for the 378 indicators in these items was 90.27%. Only one item had indicator agreement of less than 80% (Item 11. Safety practices was 79.11%). The item with the highest level of indicator agreement was Item 35. Staff professional needs, with an agreement of 97.36%. It is apparent that a high level of observer agreement at the indicator level can be obtained using the ITERS-R.

Item Reliability. Because of the nature of the scoring system, it is theoretically possible to have high indicator agreement but low agreement at the item level. Two measures of item agreement have been calculated. First, we calculated the agreement between pairs of observers within 1 point on the seven-point scale. Across the 32 child-related items, there was agreement at this level 83% of the time. For the full 39 items, agreement within 1 point was obtained in 85% of the cases. Item agreement within one point ranged from a low of 64% for Item 4. Room arrangement, to 98% for Item 38. Evaluation of staff.

A second, somewhat more conservative measure of reliability is Cohen's Kappa. This measure takes into account the difference between scores. The mean weighted Kappa for the first 32 items was .55 and for the full 39-item scale it was .58. Weighted Kappa's ranged from a low of .14 for Item 9. Diapering/toileting, to a high of .92 for Item 34. Provisions for personal needs of staff. Only two items had weighted Kappa's below .40 (Item 9. Diapering/ toileting, and Item 11. Safety practices, with a weighted Kappa of .20). In both cases the mean item score was extremely low. A characteristic of the Kappa statistic is that for items with little variability the reliability is particularly sensitive to even minor differences between observers. The authors and observers agreed that the low scores on these items accurately reflected the situation in the groups observed and that any changes to substantially increase variability would provide an inaccurate picture of the features of quality reflected in these two items. For all items with a weighted Kappa below .50 the authors examined the items carefully and made minor changes to improve the reliability of the item without changing its basic content. These changes are included in the printed version of the scale. Even using the more conservative measure of reliability, the overall results indicate a clearly acceptable level of reliability.

Overall Agreement. For the full scale, the intraclass correlation was .92 both for the full 39 items as well as for the 32 child-related items. Intraclass correlations for the seven subscales are shown in Table 1. It should be noted that the intraclass correlation for the Program Structure Subscale is calculated excluding Item 32. Provision for children with disabilities, since only a small portion of groups received a score on this item. Taken together with the high levels of agreement at the item level, the scale has clearly acceptable levels of reliability. It should be remembered that this field test used observers who had been trained and had a good grasp of the concepts used in the scale.

Table 1 Intraclass Correlations of Subscales

Subscale	Correlation
Space and Furnishings	0.73
Personal Care Routines	0.67
Listening and Talking	0.77
Activities	0.91
Interaction	0.78
Program Structure	0.87
Parents and Staff	0.92
Full Scale (Items 1-39)	0.92
All Child Items (1-32)	0.92

Internal Consistency. Finally we examined the scale for internal consistency. This is a measure of the degree to which the full scale and the subscales appear to be measuring a common concept. Overall the scale has a high level of internal consistency with a Cronbach's alpha of .93. For the child-related items, 1-32, the alpha is .92. This measure indicates a high degree of confidence that a unified concept is being measured. A second issue is the degree to which the subscales also show consistency. Table 2 shows the alphas for each subscale:

Table 2 Internal Consistency

Subscale	Alpha
Space and Furnishings	0.47
Personal Care Routines	0.56
Listening and Talking	0.79
Activities	0.79
Interaction	0.80

Program Structure	0.70
Parents and Staff	0.68
Full Scale (Items 1-39)	0.93
All Child Items (1-32)	0.92

Cronbach's alphas of .6 and higher are generally considered acceptable levels of internal consistency. Thus, caution should be taken in using the Space and Furnishings and Personal Care Routines subscales. Program Structure, Item 32. Provisions for children with disabilities was rated for only the few groups that had children with identified disabilities. The internal consistency score for this subscale was calculated excluding this item. Thus, the authors recommend using the Program Structure subscale excluding Item 32 unless most programs being assessed include children with disabilities.

Overall, the field test demonstrated a high level of interrater agreement across the scale items and at the full-scale score level. These findings are quite comparable to those found in similar studies of the original ITERS and ECERS, and the ECERS-R. All of these previous studies have been confirmed by the work of other researchers, and the scales have proven to be quite useful in a wide range of studies involving the quality of environments for young children. At the same time the scales have been shown to be user-friendly to the extent that it is possible to get observers to acceptable levels of reliability with a reasonable level of training and supervision.