Reliability and Validity of the Early Childhood Environment Rating Scale, Third Edition

As noted earlier in this document, the ECERS-3 is a revision of the widely used and documented Early Childhood Environment Rating Scale (ECERS), one in a family of instruments designed to assess the overall quality of early childhood programs. Together, these scales have been used in major research projects in the United States, as well as in a number of other countries. With only few exceptions (e.g., Sabol & Pianta, 2014), 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 ties to child development outcomes for children in classrooms with varying environmental ratings (Aboud & Hossain, 2011; Burchinal, Kainz & Cai, 2011; Burchinal, Peisner-Feinberg, Pianta, & Howes, 2002; Cryer et al., 1999; Gordon et al., 2013; Harms, Clifford, & Cryer, 2005; Helburn, 1995; Henry et al., 2004; Pinto, Pessanha, & Aguiar, 2013; Love et al., 2004; Sabol & Pianta, 2013; Sylva et al., 2004; Whitebook, Howes, & Phillips, 1989). Some of the studies show the effect of high quality as measured by the ERS instruments persists well into elementary school (Peisner-Feinberg et al., 1999), or secondary school (Sammons et al., 2011). However the relationship between overall global quality and specific child outcomes for ECERS-R, as well as other measures of child care quality, is relatively small (Burchinal et al., 2011). This new edition, ECERS-3, is designed to improve the prediction of child outcomes while maintaining the emphasis on the importance of a wide range of developmental outcomes in children.

Since the concurrent and predictive validity of the ECERS-R is well established, and the current revision maintains the basic properties of the original instrument, the focus of the first field studies of the ECERS-3 has been on the degree to which the Third Edition 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. As further studies are completed, these will be posted on the ERSI website (www.ersi.info).

After extensive revision, the authors conducted small pilot trials of the ECERS-3 in the summer of 2013, and a larger field test of the scale that autumn. The results of this field test indicated that further refinements in the ECERS-3 were needed. Subsequently, the authors completed another round of revisions in the first half of 2014 and launched a second field test in the late spring. In this second field test, a group of volunteer observers who were proficient in use of the earlier ECERS-R received training in the new ECERS-3, including field practice in which they demonstrated adequate levels of reliability. All 14 assessors attained reliability of 85%

agreement within one point on the 35 Items of the scale. Thirteen of these observers were able to attain this level of reliability with a gold standard trainer in their first two joint observations in real-life classrooms operating normally. The 14th assessor took two additional trials to get to reliability. After attaining this baseline reliability, the trained assessors were paired with one another in order to conduct the reliability study. It should be noted that these assessors were all very experienced in using the ERS instruments. One should expect a more extensive training period will be needed to train assessors new to these instruments.

The sample of classrooms in the study consisted of 50 classrooms in 4 states— Georgia (12), Louisiana (4), North Carolina (24), and Pennsylvania (10). Classrooms were recruited with a goal of having approximately 1/3 of the total be low-quality pro- grams, 1/3 be of mid-level quality, and 1/3 be of high quality, based on available data from state licensing and Quality Rating and Improvement System information. In the end, the sample is somewhat skewed, with relatively few high-scoring classrooms and more in the moderate- to low-scoring range, but adequate distribution was attained to allow for examination of use of the scale across the wide range of quality of programs available in these states. Results of the study are presented below. Assessors were rotated to the extent possible to ensure that reliability was measured across multiple assessor pairs. In each classroom two assessors rated the classroom environment independently of one another, but at the same time. The core assessment took place during a prime time of the day for exactly 3 hours, with some additional time allowed to examine the gross motor area if it was not used during the observation, and to examine materials in the classroom that were not able to be assessed during the formal observation period. In both of these added times, ratings were only allowed for the very specific Items in Gross Motor Space and Equipment, and in the Indicators related to the materials. All measures of child and teacher interactions were based on the 3-hour segment.