

Structure and Stability of the Early Childhood Environment Rating Scale¹

Richard Clifford

Introduction

To be useful to the field, a measure of environmental quality must address meaningful aspects of quality as well as be technically sound. While much has been written about the Early Childhood Environment Rating Scale (ECERS) and its companion scales, here I look at the instrument from its conceptual base and as it performs in various settings and studies. I first look briefly at the degree to which the Scale addresses meaningful aspects of quality and then go on to look at some of the properties of the Scale as it is used in early childhood programs.

A delicate balance is required of assessment instruments related to the stability of the measurement properties of the instrument. This balance is particularly important when assessing learning environments for young children. In this paper, I will examine the stability of the Early Childhood Environment Rating Scale in both its original (ECERS) and revised (ECERS-R) versions. Very simply put, to be valid and reliable, as well as useful to the field, an environmental assessment instrument must measure constructs that are relatively stable and do not change from minute to minute over the course of the day or week. However, the instrument must also be sensitive enough to be able to detect improvements or declines in environmental quality within a reasonable time period. Designing instruments that meet these general requirements and demonstrating such properties are not so simple.

The original ECERS (Harms and Clifford, 1980) was designed to provide guidance to practitioners to help them examine classroom environments in order to make improvements in the provisions for young children. To meet the demand of this kind of task, the instrument first had to be both theoretically and practically grounded. The scale had to have a strong and transparent theoretical structure. Conceptually, the items are organized into seven subscales that guide the observer to practically meaningful areas of interest in early childhood classrooms. These include 1) Personal Care Routines, 2) Furnishings and Display for Children, 3) Language-Reasoning Experiences, 4) Fine and Gross Motor Activities, 5) Creative Activities, 6) Social Development, and 7) Adult Needs in the ECERS. The ECERS-R consists of the following; 1) Space and Furnishings, 2) Personal Care Routines, 3) Language-Reasoning, 4) Activities, 5) Interaction, 6) Program Structure and 7) Parents and Staff.

Numerous studies have documented that the Scale can be used reliably by properly trained observers using this conceptual framework. While this organization is quite helpful in a practical sense, it is not empirically based. Various researchers have found an empirical structure using factor analytic techniques that is different from that represented by the seven subscales and also provides guidance in understanding learning environments for young children. These researchers have found the items of the Scales can be grouped into between two and five factors.

In this paper, I will examine data from a variety of studies using the ECERS and ECERS-R to document the internal structure and to see the degree to which it is consistent across various studies and countries. I will also detail the extent to which the instrument achieves this balance between stability and sensitivity in real world applications. Since the Scales are used internationally, I will examine the extent to which the constructs can be measured both within and across countries. I will also look at the stability of the Scales across time when using the same version of the instrument, as well as across the two versions of the Scale.

This paper addresses two specific questions about the measure:

- Is global quality as measured by the ECERS and ECERS-R (the Scales) relatively stable or is it time and situation specific?
- With the major revision of the Scale in 1998, are the original constructs of the Scale maintained? Are there systematic differences in the level of quality measured by the two versions?

While I will report on findings from several different countries, I will first focus the specific analyses for this paper on new data collected in the United States and Germany. I try to generalize from specific aspects of the samples to reach some more general conclusions on the elements of quality of early childhood care and education settings – at least as measured by the Scales.

Stability of Quality as Measured by the ECERS and ECERS-R

One can consider two kinds of stability: stability at the classroom level and at the individual child level. The ECERS and ECERS-R are designed to measure the environmental provisions made available to children in the class, but are not designed to assess for individual children the extent to which they actually make use of the various provisions. It is reasonable to expect that a child may need and receive different opportunities from day-to-day and week-to-week as the school year progresses, even though the overall availability of provisions in the classroom remains largely unchanged. In fact, it has been demonstrated many times that typical classrooms have children functioning at quite

different levels at any given point in time (see, for example, West *et al.*, 2000) and thus provisions should be made for this broad array of needs within the classroom. Thus, the global quality as measured by the Scales may be stable over time, at least to a moderate extent.

On the other hand, the needs of any individual child would be expected to change as the child develops and thus the child's actual experience with the environment may change substantially while the overall environment remains stable. To get a full picture of the impact of the environment, one would need both measures of global classroom quality as well as child specific measures of experiences with the environment. Somewhat surprisingly, a recent study found that the individual level of activities in pre-kindergarten (pre-k) classes was remarkably stable when summed across four target children in each class (Howes *et al.*, under review). Further analyses of that data may reveal individual differences. Our analyses for this paper will focus on global classroom quality.

Sometimes it is said that you cannot even measure global quality because it is always very much dependent on a given situation and is changing from moment to moment (Moss and Pence, 1994). Therefore, any single measurement will – with high probability – lead to an incomplete or even misleading assessment. However, this assertion is rarely accompanied by empirical evidence. I purport that it is wrong – at least at the level of global quality of settings as measured by the Scales. As demonstrated by the results of studies in the US and other countries, it can be seen that at the global level, quality can be measured meaningfully with confidence (Whitebook *et al.*, 1989; Peisner-Feinberg *et al.*, 2001; Zill *et al.*, 2003; Clifford *et al.*, in press). As stated above, the goal of any measure of this type is that the measure is stable enough to result in reliable estimates of classroom quality, yet capable of measuring meaningful change that occurs in classrooms over some extended period of time. I attempt to document this for the ECERS and ECERS-R.

The US data used in the following examination are from the six state study of pre-k programs conducted by the National Center for Early Development and Learning (NCELD). These data were collected during the 2001-02 school year in six US states – New York, Georgia, Kentucky, Ohio, Illinois and California. The observers completing the ratings were independent of the programs and were trained to reliability ($K \geq .60$). The settings were pre-k programs operated as part of a state pre-k initiative in each of the states, and included programs in schools and Head Start programs, as well as in non-profit and proprietary child care centers. A total of 238 classrooms (approximately 40 per state) that contained 4-year-old children (some had 3-year-olds as well) were assessed in the fall of 2001 and again in the spring of 2002. There were no systematic efforts to modify the quality of the environments in these settings during this school year. A total of 227 of

these classrooms had useable data on the ECERS-R for both the fall and spring (Clifford *et al.*, in press).

The German data are pooled from different studies conducted between 2000 and 2002 in the city of Bremen and the states of Berlin, Brandenburg, Lower Saxony, North Rhine-Westphalia, Saxony, Saxony-Anhaltinonia and Thuringia. In all studies, independent observers were trained to an observer agreement of equal or greater than 85% within one scale point. A total of 311 classrooms from these various samples had useable data on the German version of the ECERS-R called the Kindergarten-Skala Revised, here after referred to as the KES-R (Tietze, *et al.*, 1997; 2001). As with the US study, no systematic efforts to improve the environmental provisions were made in sites where we compare ratings at different points in time.

In the NCDL study, the means and standard deviations of the ECERS-R scores were 3.81 (S.D. 0.82) in the fall and 3.79 (S.D. 0.80) in the spring of the pre-k year (R. Addy, personal communication, May 25, 2004). The mean time between observations was approximately five months. The Pearson product moment correlation between individual classroom scores at these two points in time was .69 (M. Burchinal, personal communication, September 4, 2003). Because the Scales reflect a value for child-centered learning environments more often associated with early childhood programs than school environments, one might expect the quality ratings of these classrooms to change as pre-k teachers readied their children for the somewhat more formal schooling that begins at age five in US kindergartens. However, the ECERS-R scores remained remarkably stable over a relatively long period of time during the pre-k year. The NCDL study found that the ECERS-R could be meaningfully characterized as having two factors referred to as (1) Teaching and Interactions and (2) Provisions for Learning. In addition to the overall ECERS-R score being stable, the mean scores for the two factors also remained stable over the five months. Teaching and Interactions had mean scores of 4.43 (S.D. 1.29) and 4.44 (S.D. 1.22) in the fall 2001 and spring 2002 respectively with a fall-spring correlation of .60 (R. Addy, personal communication, May 25, 2004). Provisions for Learning had scores of 3.79 (S.D. 0.96) and 3.79 (S.D. 0.88), with a correlation of .72 over this time period (R. Addy, personal communication, May 25, 2004). Thus, these results point to a high stability of global quality as measured by the ECERS-R over time.

In the German sample, quality was assessed with the KES-R at two measurement points ranging from one to ten weeks apart (Tietze *et al.*, 2001). During this time, there was no intervention. Included were ten classes where the same observers applied the KES-R at the two measurement points and ten classes where different observers applied the KES-R. When the same observers were used, exact agreement of the quality assessment was reached on 73% of the items. Agreement was within one scale point on 92% of the items.

Where different observers were used, exact agreement was reached on 65% of the items and agreement within one scale point on 92% of the items. When the correlation of the KES-R total between the two points of time was calculated, the Spearman rank order correlation was .92 using the same observers and .88 using different observers. These results point to both a high test-retest-reliability and a high stability of quality across time. It may be argued that the two measurement points are only one to ten weeks apart. However, we also found the same result with the previous German version of the ECERS, the KES, when the two measurement points were eight to ten months apart (Tietze, *et al.*, 1997).

Looking at the results across studies, we find indications that the assessment of the global quality of an early childhood care and education setting as measured by the ECERS-R is stable over moderately long periods of time during a given school year where the teacher is stable in the classroom. We do not find wide variations in these provisions during this time period. On the contrary, we find indication that the quality is highly stable over time. This finding does not address whether changes can occur with intervention, simply that left to the normal progress of the program the ECERS and ECERS-R scores remain stable.

Several other studies have used the ECERS and ECERS-R to document change in environments over time. Bryant and her colleagues in North Carolina conducted a statewide evaluation of the NC Smart Start early childhood initiative over a ten year period of time (Bryant *et al.*, 2003). One of the goals of Smart Start was to raise the quality of child care across the state (Bryant *et al.*, 2003). To document changes in child care quality, the study team selected a sample of 184 centers in 18 counties which entered the program in 1994. They subsequently assessed the quality of centers in these counties multiple times between 1994 and 2002. They were able to document significant and meaningful improvement over this extended period of time using the original ECERS.

In a separate study, Whitebook and her colleagues conducted an extended study examining efforts to support child care programs in northern California working toward accreditation with the National Association for the Education of Young Children. Three levels of intervention were used with the sample of programs to help them become eligible to be accredited. They used both the original ECERS and the ECERS-R at various points in time during the project. They were able to show clear improvements in ECERS scores with dose effects dependent on the level of intervention (Whitebook *et al.*, 1997; Sakai *et al.*, 2003).

In a German study, preschool teachers were trained over a period of two years. Before and at the end of the training, the classrooms of these teachers were assessed using the ECERS-R (Erning, 2003). During this time, an increase of the ECERS-R total scores could be

observed which amounted on average to about one scale point. The NC Department of Human Resources conducted a study of the impact of training of child care teachers who were enrolled in site based college classes. They used the original ECERS to document change in the classrooms of the teachers in the classes offered by four separate institutions of higher education. They were able to document improvement in the ECERS scores, but only when there was a period of at least 90 days between pre and post assessments with intervention occurring during these times. These aforementioned studies verify the ability of the ECERS and ECERS-R to detect change in environmental conditions over a moderate to long intervention.

In summary, the assessment of global quality as measured by the Scales seems not to be affected by the specific situation on the day of the assessment. Rather, this quality seems to be relatively stable over time. At the same time, the ratings are sensitive enough to detect changes that occur as the result of improvements.

Differences in the Mean Quality Levels between ECERS and ECERS-R

In the transition from ECERS to ECERS-R, not only items have been modified and added, but the rating scales themselves have been methodologically changed. In the original version of the ECERS, the rater had more freedom in assigning a value compared to the strict indicator system used in the ECERS-R. When comparing quality ratings done with ECERS and ECERS-R it is important to know if differences are due to real quality differences or to the differences in applying the rating scales. So far, researchers have found mixed results (see Table 1).

Table 1: Comparison of ECERS and ECERS-R Totals

| | ECERS | | | ECERS-R | | | Difference in means |
|--|--------------|------|-----|--------------|------|-----|---------------------|
| | No. of items | Mean | SD | No. of items | Mean | SD | |
| US | | | | | | | |
| a. same sample, different observers: n=68 classroom (ration) | 37 | 4.91 | 0.7 | 43 | 4.87 | 0.8 | .004 |
| Germany | | | | | | | |
| a. different samples, for KES n=103 classrooms, for KES-R n=180 classrooms (Tietze et al., 2001) | 29 | 4.51 | 0.7 | 39 | 4.06 | 0.8 | 0.45 |
| b. same sample, n=159 classroom, same rater for KES and KES-R | 29 | 4.77 | 0.8 | 39 | 4.19 | 0.6 | 0.58 |

In a study conducted in the US in 68 classrooms (Sakai *et al.*, 2003), two trained raters independently observed the same classroom at the same measurement point, one with the ECERS and the other with the ECERS-R. For the ECERS, a mean of 4.91 was found. The mean for the ECERS-R was 4.87 and, thus, almost identical. The standard deviations are also comparable. According to this study, the mean is not affected by the changes in the scales.

A different result was found in Germany. In a first study, means and standard deviations of 103 classrooms measured with the ECERS and 180 different classrooms measured with the ECERS-R were compared. The ECERS total mean was about half a scale point higher than the ECERS-R total mean, whereas the standard deviations were about the same. This mean difference might be due to the more strict indicator system in the ECERS-R or to a different quality level in the two samples. In a second study of 159 classrooms, the same observers used ECERS and ECERS-R at the same measurement point. While using the same observers might lead to biased ratings, again the ECERS total mean was about half a scale point higher than the ECERS-R mean, whereas the variance was not affected.

Results from these three studies do not provide a conclusive answer as to whether or not quality ratings obtained with the ECERS and the ECERS-R are comparable. Whereas in the US, no mean differences between ECERS and ECERS-R mean scores are found, the more strict rating modus of the ECERS-R seems to systematically decrease the quality assessment by about half a scale point in Germany. This has to be taken into account when comparing quality ratings done with ECERS and with ECERS-R. Interestingly – this is not included in Table 1 – there are some indications that in Germany, the decrease seems to be different for the factors Teaching and Interaction and Space and Materials. In the first German study, the decrease for Space and Materials amounts to 0.8 scale point, whereas the quality rises for one tenth of a scale point for Teaching and Interaction. Thus, it seems to be that more easily observable aspects of the Space and Material dimension are more affected by the transition to the ECERS-R than aspects related to teacher-child interactions. However, this result was found only in Germany.

Summary

This paper has discussed some questions that are of high importance for the use of the ECERS and the ECERS-R in theoretical and research perspectives, as well as in regard to practical improvements. An effective instrument for measuring the global quality of an early childhood setting has to be stable over time as well as sensitive to introduced changes. That is, the overall quality rating should not be affected by the specific situation on the given day of assessment when no specific reasons to expect changes have occurred, such as a substantial change of the teachers, the program or quality improvement measures. This does not mean that individual children do not experience

changes in the stimulation they experience according to their changing needs (like improving capacities with age). However, it is hypothesized that such changes on the environments experienced at the individual level can occur or can be embedded in rather stable overall quality at the setting level. To get a full picture of the impact of the environment, one would need both measures of the global classroom quality as well as child specific measures of experience with the environment, such as the Emerging Academics Snapshot (Ritchie *et al.*, 2001), which is used to assess children's experiences in the classroom (i.e., their engagement with activities and interactions with adults). The ECERS and the ECERS-R are related to the global classroom quality. I have shown that at this level both requirements of a good measurement instrument are fulfilled. Quality as measured by the ECERS and the ECERS-R is highly stable across moderate time intervals up to a year. At the same time, studies show that the Scales are also sensitive to changes occurring after quality improvement measures have been implemented.

Quality of an early childhood setting is not conceived as an undifferentiated construct. Rather, we assume that different areas or dimensions of quality exist. In designing the ECERS and the ECERS-R, a structure of seven subscales was assumed. While both the original and the revised instruments are divided into these subscales, these subscales were developed for ease of use and specifically designed for practitioners and, thus, may not represent empirically separate dimensions of the environment. In fact, there has been some debate as to whether the global environmental quality can be broken down into more discrete factors using the ECERS or similar instruments. For a good quality measure we expect a meaningful empirical factor structure which is not affected by the transition from the ECERS to the ECERS-R and which is not related to the characteristics of a given sample but is rather independent of specific samples and – at least to some degree – of the given situations of different countries.

One change in the transition from the ECERS to the ECERS-R is the use of a more strict indicator system supporting the ratings. Even when in the original version of the ECERS, descriptions were given for the ratings of 1, 3, 5 and 7, the rater had more freedom in assigning a value compared to the indicator system used in the ECERS-R. The question is whether or not this change leads to a decrease in the mean quality level assessed, i.e., if differences between ratings with the ECERS and the ECERS-R are due to real quality differences or to the differences in applying the rating scales. For this question, we find mixed results. The ECERS-R ratings may lead to a systematic decrease of the assessed quality level by about half a scale point. However, this was only found in German samples, whereas in the US the means seem not to be affected by the different methodological features of the rating scales.

References

- Bryant, D., Maxwell, K., Taylor, K., Poe, M., Peisner-Feinberg, E. and Bernier, K. (2003). *Smart Start and Preschool Child Care Quality in NC: Change over Time and relation to Children's Readiness*. Chapel Hill, NC: FPG Child Development Institute.
- Clifford, R., Barbarin, O., Chang, F., Early, D., Bryant, D., Howes, C. *et al.* (in press). *What is Pre-kindergarten? Trends in the Development of a Public System of Pre-kindergarten Services*.
- Erning, G. (2003). *Qualitaetsentwicklung in Kindergaerten. Abschlussbericht der Fortbildung 2001-2003 (Quality Improvement in Preschools)*. Unpublished paper. Germany: University of Bamberg.
- Harms, T. and Clifford, R. (1980). *Early Childhood Environment Rating Scale*. New York: Teachers College Press.
- Howes, C., Ritchie, S., Burchinal, M., Bryant, D., Early, D., Clifford, R., *et al.*, (under review). *Practices in Pre-kindergarten Programs: Pre-academic Activities, Teacher Behavior and Instructional Structures*.
- Moss, P. and Pence, A. (Eds.) (1994). *Valuing Quality in Early Childhood Services*. New York: Teachers College Press.
- Peisner-Feinberg, E., Burchinal, M., Clifford R., Culkin, M., Howes, C., Kagan, S. and Yazejian, N. (2001). The Relation of Preschool Child-care Quality to Children's Cognitive and Social Developmental Trajectories through Second Grade. *Child Development*, Volume 72, No. 5, pp. 1534-1553.
- Ritchie, S., Howes, C., Kraft-Sayre, M. and Weiser, B. (2001). *Emerging Academic Snapshot*. Unpublished measure, University of California at Los Angeles.
- Sakai, L., Whitebook, M., Wishard, A. and Howes, C. (2003). Evaluating the Early Childhood Environment Rating Scale (ECERS): Assessing Differences between the First and Revised Edition. *Early Childhood Research Quarterly*, Volume 18, pp. 427-445.
- Tietze, W., Schuster, K. and Rossbach, H. (1997). *Kindergarten-Einschaetz-Skala* (German Version of the Early Childhood Environment Rating Scale by Thelma Harms/Richard M. Clifford). Neuwied: Luchterhand.

Tietze, W., Schuster, K., Grenner, K. and Rossbach, H. (2001). *Kindergarten-Skala: revidierte Fassung (KES-R)* (German Version of the Early Childhood Environment Rating Scale Revised Edition von Thelma Harms/ Richard M. Clifford/ Debby Cryer). Neiwied: Luchterhand.

West, J., Denton, K. and Germino-Hausken, E. (2000). *America's Kindergarteners*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Whitebook, M., Howes, C. and Phillips, D. (1989). *Who Cares? Child Care Teachers and the Quality of Care in America. The National Child Care Staffing Study*. Oakland, CA: Child Care Employee Project.

Whitebook, M., Sakai, L. and Howes, C. (1997). *NAEYC Accreditation as a Strategy for Improving Child Care Quality: An Assessment by the National Center for the Early Childhood Work Force*. Washington, DC: NCECW.

Zill, N., Resnick, G., Kim, K., O'Donnell, K., Sorongon, A., McKey, R., et al. (2003). *Head Start FACES 2000: A Whole-child Perspective on Program Performance, Fourth Progress Report*. Washington, DC: Administration on Children, Youth, and Families, U.S. Department of Health and Human Services.

Note:

- 1 This paper is based on a joint paper produced by Richard Clifford and Hans-Guenther Rossbach, University of Bamberg, Germany: Clifford, R. and Rossbach, H. (in press). Structure and Stability of the Early Childhood Environment Rating Scale, (in) Cryer, D. (Ed.). *A World of Improvement: Promoting Quality Early Childhood Programs for All Children*.