MEASURING QUALITY IN PRE-KINDERGARTEN CLASSROOMS: ASSESSING THE EARLY CHILDHOOD ENVIRONMENT RATING SCALE

By

Kerry G. Hofer

Dissertation

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Teaching and Learning

May 2008

Nashville, Tennessee

Approved:

Professor Dale Farran

Professor David Cordray

Professor David Dickinson

Professor Mary Louise Hemmeter

To my parents, Buddy and Sharon, my biggest cheerleaders;

To my brothers, Kevin and Kyle, my comic relief;

And to my sweet husband, Mark, my lifeline and my heart.

ACKNOWLEDGEMENTS

This research would not have been possible without funding from the Institute for Education Sciences Experimental Education Research Training (ExpERT) program (R305b040110). This predoctoral fellowship afforded me the unique opportunity to participate in a program that combined both content in early education and sound quantitative research methodology. In addition, I owe much thanks to the field experts who graciously agreed to contribute to my dissertation research. Their assistance added great value to the study.

All of the members of my committee, Dale Farran, David Cordray, David Dickinson, and M.L. Hemmeter, were invaluable to me in the dissertation process. Each brought a unique set of knowledge and skills to the committee, which they shared graciously in order to assist me on this project as well as to further the field of early education. I believe it is a rare experience for a doctoral student to have committee members that not only possess knowledge critical to the research but also demonstrate such commitment to and care for the student, and I was lucky enough to have such a committee.

I would especially like to thank Dale Farran, my advisor and dissertation committee chair. In addition to mentoring me in my program path, facilitating my growth as a researcher, and providing stellar research opportunities with which to be involved, Dr. Farran remained a strong advocate for my interests and needs as a student. She went above the call of an advisor, working not only to ensure that I had the best possible doctoral experience but providing incredible opportunities that have better prepared me for my professional career. I have learned much from her. I also owe thanks to Dave Cordray, director of the ExpERT program and member of my dissertation committee. His contributions to this dissertation were critical, and I

could not have written it without his help. Drs. Farran and Cordray have been supporting me from the very beginning of my Vanderbilt experience, and I am extremely grateful for having known and been mentored by them.

I am particularly grateful for the students and staff that I have worked with during my time at Vanderbilt in the *Focus on Early Learning* group. I feel blessed to have had the opportunity to work with such a talented, committed, and fun group of people. I owe special thanks to Sarah Shufelt and Mackenzie Richardson, who helped with much of the data entry for this project. In addition, I am especially grateful for Betsy Watson, with whom I trudged through the dissertation process. Her encouragement was invaluable, and I would not have wanted to go through the experience with anyone else.

I could not have survived the past several years without my family, my friends, and my faith. I know that God led me to Vanderbilt despite my misgivings about returning to the life of a student, and I know that, as always, following Him was the right decision. Not only has He stayed with me, but He helped me to write on those days when my dissertation weighed too heavily on my mind for words to come. I am incredibly grateful to my parents, who have seen me through every stage of my life with constant encouragement and support. Somehow they always know just when I need it. Throughout my years at Vanderbilt, I have also received continual and enthusiastic support from my brothers, my in-laws, my extended family, and my friends. I am blessed to be able to say that there are too many of them to name. I owe a huge debt of gratitude to my husband, Mark, who stood by me even on my most trying days and pushed me to do what I often thought I could not. He knew the challenges that awaited us in Nashville better than I did but always put my needs above his own, and every day I am amazed by his love. Any success that I may have is owed to these people who love me the most.

TABLE OF CONTENTS

Page

Chapter

Appendix

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER I

INTRODUCTION

Statement of the Problem

Over half of the children between the ages of three and five in the United States are enrolled in some form of preschool education program (U. S. Department of Education, National Center for Education Statistics, 2006), and the number continues to rise. Today's preschoolers will soon begin their formal schooling years. In another 15 years, most of those children will be entering the country's work force. As the stepping stone into the world of education that will be responsible for preparing those students for their lives as public citizens, the preschool arena serves an important purpose.

One of the reasons that the number of children enrolled in preschool programs is increasing concerns the growing number of families with two working parents and the number of single-parent families. With all available parental figures out of the house during the day, parents turn to childcare. However, simply having places for children to be while their parents are at work is not the only requirement for such an expanding population. Of greater import is providing places that offer quality experiences for the children enrolled in them. The term *quality*, however, represents a fairly general and subjective concept, particularly in education. Quality assurance in other fields can often be strictly defined. For instance, in consumer goods, the quality of a product might be thought of in terms of its cost to the buyer, its shelf life, or its stability. But when the "good" is an intangible such as an experience, the definition of quality becomes more elusive.

A child’s experience in a classroom is made up of different types of interactions: interactions with the caregiver(s), interactions with other children, interactions with the physical space, interactions with available materials, etc. Such interactions can be of a social, academic, behavioral, or routine nature. Those interactions might be influenced by more distal characteristics like teacher wage, teacher education requirements, and program support. As a result, defining the quality of a child’s interactions within the preschool environment can be quite difficult.

There are many easily-measured characteristics that can be objectively observed that some may think of as representing quality in classrooms or programs. For instance, the number and type of materials in a classroom, the ratio of teachers to children, and the educational attainment of the staff are often used as components of a classroom’s quality assessment. But such characteristics are only indirect measures of quality through which high-quality interactions between teachers and children may or may not occur. Alternatively, one can look directly at those interactive behaviors inside of the classroom, to develop a picture of classroom quality, but this alternate method can often be more time consuming, as well as more subjective. Moreover, interactions may differ by child in the classroom.

Despite the difficulty in explicitly defining the components of quality in preschool, the *idea* of quality is widely regarded as a critical element of a young child’s first experiences in education. In a 1998 speech, former Vice President Al Gore remarked, “Quality child care isn't a luxury, it's a necessity. It not only gives parents peace of mind -- it gives children safe places to learn and to grow.” Thus, ten years ago, the idea that quality preschool experiences should be available to all children, regardless of family income, was recognized, as well as the idea that parents, regardless of family income, had the right to seek out and find quality programs.

Parents are not the only group concerned with the quality of their child’s experiences. In his book, *The Sandbox Investment*, University of California’s Public Policy Professor David L. Kirp wrote:

> The age-old parental desire to give one’s own kids the best chance to succeed has evolved into a nationwide push for high-quality preschool that, like K-12 public education is paid for with tax dollars and open to all. Nor is it just parents who are behind this effort. The big-tent coalition of pre-K supporters includes politicians and pedagogues, philanthropists, pediatricians, and police chiefs. (2007, p.3)

The specific definition of quality is somewhat hard to pin down, and parents, politicians, and advocates may be using the same term with quite different things in mind. Nevertheless, they all agree on the importance of quality.

Because quality is so highly valued and has become an increasingly present topic in discussions of early childhood education, researchers, practitioners, policy makers, and parents have looked to different instruments purported to assess the quality of a program. An important first step for a developer in designing an instrument to measure quality is deciding what his or her definition of quality is. Without a theory behind the measure, it would be impossible for the developer to assign high and lower ratings to each aspect being measured. A developer’s quality definition can often give the tool’s user some much-needed guidance into interpreting the results of the assessment. Unfortunately, the definitions held by developers are often quite implicit, leaving the responsibility in the hands of the user to glean the developer’s perspective from the aspects included in the tool itself.

Preschool programs across the country are evaluated for their quality, and those same evaluations are used by community members for different purposes. Scores presented as indications of the quality of a program are often made available to parents in order to aid them in the selection of their child’s preschool placement. These same types of scores have been used by researchers to evaluate the influence of a child’s early care environment on children’s educational success. Quality scores are also used by the policy world in decisions about program funding and childcare reform.

It is this last use of quality assessments that is of the most concern. Currently child care quality assessment instruments are being used to determine the amount of money that is awarded to preschool programs, effectively tying a program’s score to, among other things, the salary of the program’s staff. When the livelihood of citizens is determined in part by an instrument’s score, such an instrument must be critically examined. What program characteristics the instrument assesses, how those characteristics are discerned, and how the instrument’s final score is determined and interpreted are all areas that should be considered carefully.

The current use of quality assessments has become one in which researchers are exporting tools designed for research and evaluation purposes into the realm of policy, a realm in which real consequences for the people involved are tied to the resulting assessment scores. The properties of such assessment tools must be understood in order for their use in evaluation, and hence their tie to public sector consequences, to be valid.

A good assessment tool must incorporate the properties often discussed in classic test construction literature, an even more important requirement if the results of a test are tied to policy decisions. For example, the instrument should have demonstrated at least the possibility of strong inter-rater reliability at the level that quality and often policy decisions are made. In addition, the levels of a scale should be consistent in their hierarchy; the components of an item should reflect the same construct that the item as a whole reflects, and the items in a subscale should reflect the same construct that the subscale as a whole represents. If multiple methods of scoring the same instrument are possible, those alternative methods should not yield conflicting outcomes, especially important when those results lead to policy decisions about the funding of a site. The results of an instrument should be relatively stable across time. Though teachers and classrooms may change over the course of a year, it is important to consider whether the time of year that a classroom is measured might affect their eventual funding. It may also be important to early childhood educators that the scale reflects the prevailing wisdom of the field of what constitutes quality in an early educational environment. This study sought to examine the properties listed above within the currently most widely-used instrument designed to measure quality in early child care.

Objectives

The objectives of this study were to critically examine the Early Childhood Environment Rating Scale-Revised Edition (ECERS-R; Harms, Clifford, & Cryer, 1998) using four hierarchical perspectives. First, the study examined how different scoring methods applied to the same ECERS-R data might influence the final quality ratings that a classroom achieves. Second, the study examined how the ECERS-R reflects the current view in the field about what aspects of a classroom contribute to quality. Third, the psychometric properties the ECERS-R and a new version of the instrument based on items important to field experts were examined. And fourth, this study sought to examine policy issues associated with a new version of the ECERS-R based on expert opinion. This research sought to provide the realms of education, research, and policy with a more detailed picture of the meaning behind the assessment tool's scores, as well as to suggest ways that this measure might be improved.

CHAPTER II

REVIEW OF THE LITERATURE

Historical Perspective on Early Childhood Education

Two Strands in Early Childhood History

Historically, the practice of caring for young children has run the gamut from entirely care-based to almost completely academically-based, and aspects of both perspectives can still be seen in the current state of preschool education. From its origination, early childcare in the United States began with two separate sets of practices. The following brief look into the history of child care in America is taken primarily from Kamerman and Gatenio (2003). The first foundation for early childcare began in the 1830's with child care centers or day nurseries. These programs were designed specifically to provide care for the children of working mothers and were primarily supervisory, focusing on providing children with the basic custodial care they required. Programs that focused more on educating young children – the second foundation – began with early educational programs and kindergartens, such as those begun by Freidrich Froebel, with the development of nursery schools (Wolfe, 2002). Historical educational theorists like John Dewey and Friedrich Froebel portrayed early education as providing children with unique opportunities to develop their academic skills. In his book, *The Education of Man*, Froebel stated that, "To lead children early to think, this I consider the first and foremost object of child-training" (as cited in Wolfe, p. 77). There was a boost in the number of nursery schools during the 1920’s as middle-class parents sought educationally enhancing experiences for their children.

During the 1960’s and 70’s, both child care centers and nursery schools grew in number. President Johnson’s War on Poverty speech, which led Congress to pass the Economic Opportunity Act, focused citizens’ attention on the experiences of children from impoverished homes, which in turn led to the compensatory education movement (Kamerman & Gatenio, 2003). This movement conceptualized children with minority status and high poverty families as being at a disadvantage when compared to middle-class children in school. Therefore, those disadvantaged children needed extra assistance in order to “compensate” for their disparate backgrounds. People were beginning to realize both the importance of children’s early experiences and the need for opportunities to participate in similar experiences for all young children. Researchers lauded early education programs that prepared children for school as well as provided them with the health and nutrition that children from poor families were not getting in their homes (Kamerman & Gatenio). A belief in compensatory education led to the formation of the Head Start Program in 1965. For middle-class families, increases in women participating in the labor force added steam to the push for quality care outside of the home. At the same time, parents from middle-class homes begin to view preschool education as not only valuable for their young children but essential for facilitating their transition into formal schooling.

The two strands of early education, academic-focused and care-focused, were pretty easily delineated in the past. Since the 1990’s, however, there has been some movement to integrate the objectives of education and care, and the line between care-focused and academic-focused programs today is becoming less and less clear (Melhuish, 2001). As preschool has made its way into the public school system, there has been some overlap of the two strands of thought. Public preschool, perhaps because of its integration into the academically focused environment of formal school, has adopted an increasingly educational concentration. At the same time, however, due to the age, development, and mixed socioeconomic status of its students, public preschool is sometimes held accountable for its care-focused programming. Outside of public education, there exists a wide variety of programs for young children with great differences in orientation toward care or academics. This variety is exacerbated by the "categorical funding" of such programs, an issue discussed later in this paper (Kamerman & Gatenio, 2003).

Current Status of Early Childhood Programs

As the debate continues about which of the competing goals (academics or care) should predominate in early childcare programs, the number of children enrolled in preschool or daycare programs continues to rise. This is due in part to the increasing number of women in the workforce. As more mothers take on jobs outside of the home, the need for care facilities for their children becomes greater. In 1940, 28% of all women in America were in the labor force, and that number rose to 60% by 1997 (Smith & Bachu, 1999). In addition, with the increasing rate of divorce and single parent homes, more children are being raised in households with only one parent, which, when that parent must work, also necessitates care facilities for children to attend during the day. In 1960, 9% of children under the age of 18 in the United States lived in single-parent homes, and by 1999, that percentage had risen to 27% (Sado & Bayer, 2001).

The number of children currently enrolled in preschool programs is larger than ever, leading to an even greater variety of programs to serve them. As a result, the focus on assessing the quality of those preschool environments has intensified. The experiences that preschool classrooms provide for children cannot be ignored in a country that has no publicly funded early childhood care system but has over half of its 3- and 4-year old children taking part in some form of educational experience prior to formal schooling (See Figure 1, U.S. Census Bureau, n.d.).

An increasing focus on early childhood education led to funding of longitudinal studies of the effects of preschool interventions such as the Abecedarian Project (The Carolina Abecedarian Project, n.d.), an endeavor begun in the early 1970's. Such studies claimed that exposure to early childhood experiences led to better outcomes in formal schooling. Results of such studies have been used to justify the need for further funding for early childcare programs, independent of quality specifications.

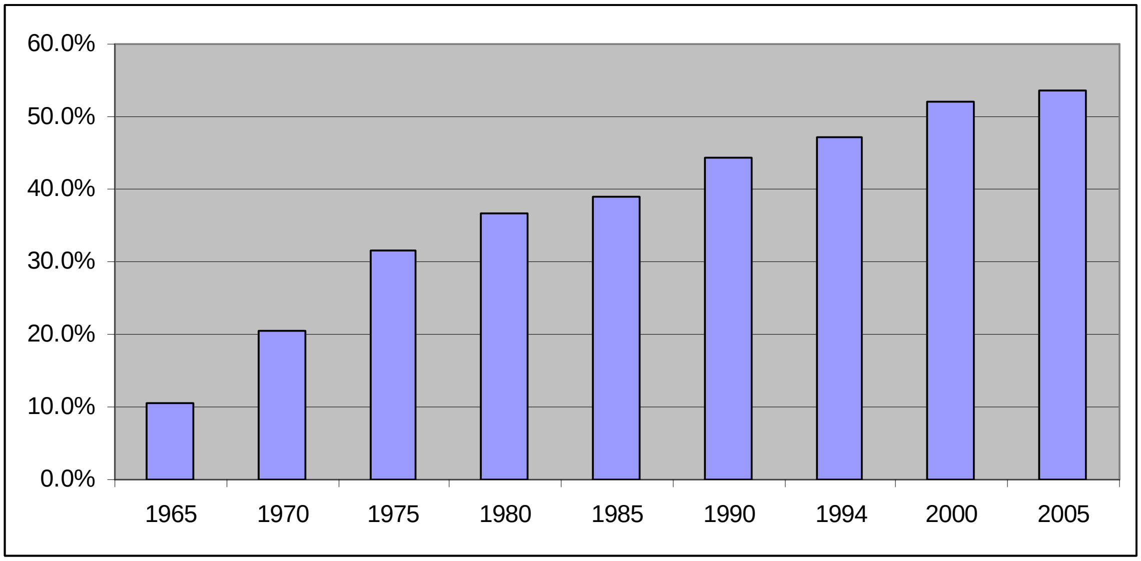


Figure 1. Percentage of American 3- and 4-year olds enrolled in school across time ("school" includes any type of public, parochial, or other private school including nursery schools, kindergartens, etc.).

Quality Variation. A major contribution of research on early childhood educational settings has been the demonstration of the degree to which quality differs within states, across states, and across countries. Even when quality is defined and assessed in the same way across settings, it is clear that quality can vary widely within and between those settings. The European Child Care and Education Study (Tietze, Cryer, Bairrão, Palacios, & Wetzel, 1996) observed early childcare across nations including the United States, Germany, Austria, Spain, and Portugal. When quality was measured with the same instruments in all participating countries, the United States had the widest range of quality scores compared to all other countries in the study. This finding could relate to state differences in preschool eligibility requirements, requirements for the number of children served, the amount of early education expenditures, and preschool program standards, all found to vary widely across American states (Bryant et al., 2002). For instance, only 28 out of the 38 states with pre-kindergarten initiatives require lead teachers to have training specific to preschool education, and only 15 of those 28 states require teachers to obtain a bachelor's degree (The National Institute for Early Education Research, 2006). Quality is not defined at a national level and therefore varies quite significantly from one location to the next.

A significant problem often affecting uniform quality assurance in early childhood is that the child care options available to parents today are spread across several broad categories, each with unique sources of funding. The federal support for early childcare is a "patchwork of programs and funding streams, rather than a seamless system of early care and education" (Schaefer, 2003, p.1). According to Schaefer, there are six main funding sources at the federal level for early childcare programs: Child Care and Development Block Grant, Head Start, Temporary Assistance to Needy Families, Social Services Block Grant, Child and Adult Care Food Program, and Dependent Care Tax Credit. Although the federal government provides some monetary support for early childhood programs, decisions about quality regulations and standards often do not come out of federal law but rather state-level regulations (with the exception of specific requirements set by specific funding sources). Given the various funding sources which all have their own goals and values associated with early childhood programming, one cannot begin a discussion of quality in early childhood without acknowledging the difficulty in coordinating efforts to measure, maintain, and improve it.

Defining Quality in Early Childhood Education

When discussing quality early education, several terms are often used without providing clear definitions and ensuring common understanding. In the general literature, quality is depicted in terms relating to excellence, value, conformance to specifications, and/or meeting customer expectations (Reeves & Bednar, 1994). The reasons parents choose a preschool might not include quality as a priority at all. For example, in a study of parental beliefs about their children's education, a larger percentage of parents reported their choice of preschool involved the services and facilities of the school than parents who reported their decision was based on the academic emphasis, curriculum, social skills focus, self-esteem focus, or teacher warmth (Stipek, Milburn, Clements, & Daniels, 1992). In addition, parents from a specific culture might value certain characteristics of education that their culture would consider to comprise quality that might be different from another subgroup. For example, differences in teaching practices and strategies articulated by teachers have been found in research examining instruction in early childhood programs led by teachers of different cultural backgrounds, including African-American, White, and Latino (Wishard, Shivers, Howes, & Ritchie, 2003). When the definitions of quality are not consistent, general statements about the relationship between quality and other variables are made more difficult.

Alternative Foci

Structural Features. One strand of research on quality that was particularly strong in the past focuses on what are often referred to as structural characteristics of early childhood environments. *Structural characteristics* involve those variables associated with a school environment that are affected by outside forces, often government regulations. As a result, structural variables are also often referred to as *regulatable factors*. Commonly used structural variables in quality analyses are teacher wages, class size, teacher education levels or experience, and the ratio of teachers to children in the classroom. *Structural quality,* therefore, defines the quality of the environment in terms of these types of characteristics that are more distal in relation to the classroom itself.

Issues related to the health and safety of an environment could also be considered structural factors because they are regulated by child care licensing standards. Such standards ensure that an environment provides safe playground equipment for the children, accessible fire extinguishers, posted emergency exit plans, hygienic food preparation and diapering methods, etc. However, health and safety concerns also appear as process quality characteristics, described in the next section.

Process Features. Researchers often examine the relationship between the structural or regulatable characteristics of a program and the process quality of a classroom. *Process quality* refers to the quality of a child's direct experience in the childcare environment, such as the nature of a child's interactions with teachers, peers, or materials in the classroom. Also often included under this category are aspects concerning the materials available to the children and the health and safety of the environment. Aspects such as interactions with materials and engagement during lessons are called *process factors*, or *dynamic factors*. Despite the fact that the term *process quality* is often used in the literature on quality, some lines of research emphasize different components of classroom processes than others. Generally, process quality involves a child's direct interactions with the environment, but certain aspects must be in place for those interactions to occur, which broadens the definition of process quality to interactions and the environment as opposed to only interactions with the environment. Because process quality can be used to refer to a wide range of child experiences, with some researchers referring to the physical structure (e.g., room arrangement, materials present) of the environment while others describe interactions, it is difficult to generalize across studies claiming that process quality is important for children's development when there is a lack of consistency in what process quality is across those studies. The terms *structural factor*, *process factor*, *structural quality*, *process quality*, *dynamic factor,* and *regulatable factor* are frequently used in the literature discussing the quality of early childhood settings, and it is difficult to understand the author's intentions if the definitions of such terms are not clear. In the following discussion of quality and its relationship to other variables, special attention is paid to pinpointing the researchers' definitions of quality.

Representative Classroom Observational Measures

Variations in definitions of quality. Not only have people focused on different aspects of process quality, but researchers have also disagreed about whether structural or process variables should be given the most attention. Scarr, Eisenberg, and Deater-Deckard (1994) suggested that the measurement should be tailored to suit the purpose of the quality assessment. For example, extensive checklists of quality components may fit one center's purpose of attempting to understand the quality of their classroom environments at a detailed level, while more general scale measures may suit the purpose of a community organization that wishes to compare centers in its area.

Several commonly-used measures of quality include lists of specific components and require long observation periods in classrooms to determine which of those components are present. For example, the Early Childhood Environment Rating Scale-Revised Edition (ECERS-R; Harms, Clifford, & Cryer, 1998) consists of 470 individual indicators to be checked off during one observation. With so many individual indicators of quality that can be summed and averaged into various ways of representing a classroom's quality, an individual center can gain very detailed information about their classrooms. In contrast, policy makers that might have a more comparative purpose, looking more broadly at quality across an entire district or state to get a sense of the range of general quality represented by a group of programs, might necessitate a measure that can be used more efficiently and cheaply, such as quick tallies of teacher-child ratios or teacher education levels (more structural variables).

There has been a dramatic increase in the use of program review tools, standards, and/or observational scales designed to examine early childcare quality since the 1970's (Lee & Walsh, 2004). Numerous instruments have been developed to assess quality as viewed through different lenses, some stressing process quality and others focusing on structural quality, while others look at a combination of both. Each instrument examines particular characteristics of an environment thought to be important to the overall quality of the classroom. Differing perspectives on the meaning of quality should not lead one to conclude that quality cannot be measured accurately, but one must consider the beliefs of the developers of an instrument when determining the implications of its quality assessment (Melhuish, 2001). A research report finding a relationship between quality and child literacy competency might be taken at first glance by center directors to mean that an increase in scores on their own quality assessment tool, such as the ECERS-R, will result in better child literacy scores. However, if the researchers had used a measure of structural quality rather than process quality, the true findings might be simply that child care centers that paid their teachers more tended to have children with higher literacy outcome scores than in centers that did not pay their teachers as highly. This is often the case; child care centers who can afford to pay their teachers well typically enroll children from affluent families who perform better on measures of academic progress due to factors involving the home literacy environment, access to print, rich language environments, etc. Therefore, such a study would have limited implications for increasing quality in child care centers. It is important that the quality variables focused on by researchers are considered when determining what the true relevance of study findings are.

There are a number of instruments designed to assess early educational environments, and several of these instruments have been shown to measure similar aspects of the classroom. Table 1 lists several popular instruments that purport to assess quality; all use observational data. These instruments focus primarily on the process aspects of quality, differing in their concentration either on the classroom environment as a whole, including physical characteristics, or on instruction and interactions alone. Though each has a different focus, many of the instruments are consistently used in the assessment of early childhood classrooms both domestically and internationally.   One instrument designed to assess quality is the Early Childhood Environment Rating Scale (ECERS, Harms & Clifford, 1980), perhaps the most widely-used measure to evaluate program quality (Sakai, Whitebook, Wishard, & Howes, 2003). According to the developers' website, the ECERS was developed to investigate process quality in early childcare environments (*Environment Rating Scales*, n.d).

Table 1

Quality Assessment Instruments

Instrument	Developers
Focus on Classroom Setting:	
Early Childhood Environment Rating Scale (ECERS)	Harms & Clifford, 1980
Early Childhood Environment Rating Scale-Revised (ECERS-R)	Harms, Clifford, & Cryer, 1988
Infant/Toddler Environment Rating Scale (ITERS)	Harms, Cryer, & Clifford, 1990
Quality of Daycare Environment (QDCE) Morgan, & Rock, 1996	Bradley,Caldwell, Fitzgerald,
Daycare Quality Assessment Inventory (DQAI)	Peterson & Peterson, 1986
Early Childhood Classroom Observation Scale (ECCOS)	Bredekamp, 1986
Assessment Profile for Early Childhood Programs	Abbott-Shim & Sibley, 1987
Childcare Facility Schedule	World Health Organisation, 1990
Focus more on Interactions and Instruction:	
Classroom Assessment Scoring System (CLASS)	LaParo & Pianta, 2003
Early Childhood Environment Rating Scale-Extension (ECERS-E)	Sylva, Siraj-Blatchford,& Taggart, 2003
Snapshot	Ritchie, Howes, Kraft-Sayre, & Weiser, 2002
Early Childhood Classroom Observation Measure (ECCOM)	Stipek, 1996
Classroom Observation System for Kindergarten (COS-K)	National Center for Early Development & Learning, 1997
Caregiver Instruction Scale (CIS)	Arnett, 1989
Daycare Environmental Inventory (DCEI)	Prescott, Kritchevsky, & Jones, 1972
Adult Involvement Scale (AIS)	Howes & Stewart, 1987
Observational Record of the Caregiving Environment (ORCE)	NICHD ECCRN, 1996
Classroom Practices Inventory	Hyson, Hirsh-Pasek, & Rescorla, 1990
Early Childhood Observation Form (ECOF)	Stipek, Daniels, Galuzzo, & Milburn, 1992
Early Language and Literacy Classroom Observation (ELLCO)	Smith & Dickinson, 2002

Intercorrelations among observational measures. Because of its extensive use, the ECERS has become somewhat of an anchor scale for other instruments to be developed. Many research studies utilizing other instruments to assess quality have attempted to correlate those instruments with the ECERS. Scores from the ECERS have been correlated with the Classroom

Assessment Scoring System (CLASS; LaParo & Pianta, 2003) and the Snapshot (Ritchie, Howes, Kraft-Sayre, & Weiser, 2002) in several research studies (LaParo, Pianta, & Stuhlman, 2004; Pianta et al., 2005). Additionally, the ECERS has been correlated with the Caregiver Instruction Scale (CIS; Arnett, 1989), the Early Childhood Observation Form (ECOF; Stipek, Daniels, Galuzzo, & Milburn, 1992), and the Adult Involvement Scale (AIS; Howes & Stewart, 1987) in other studies (Peisner-Feinberg et al., 2001; Wishard et al., 2003). Researchers have also found correlations between the ECERS and the Assessment Profile for Early Childhood Programs (Abbott-Shim & Sibley, 1987), a similar measure examining the quality of the classroom setting (Phillips, Mekos, Scarr, McCartney, & Abbott-Shim, 2000; Scarr et al., 1994). Although some of these instruments relate more strongly to the ECERS than others, researchers have found statistically significant correlations between the ECERS and all of the above-mentioned measures. Because of the widespread use of the ECERS, other instruments that have been shown to correlate with the ECERS are more likely to be used for research, policy, and/or practice purposes than instruments without such correlational displays.

Variations in Quality

The extent to which variations in the measurement quality are important is perhaps best assessed through their relationship to child outcomes. Aware of this need, researchers who have developed assessment instruments have attempted to show relationships between their measure of quality and a variety of outcomes thought to be beneficial for children. The majority of this research has examined the correlation between measures of classroom quality and what are *immediate* outcomes. An immediate outcome is a variable that is measured within the same time frame as the quality assessment was measured. Sometimes, this second presumed dependent variable is actually measured at the exact same time as the measure of classroom quality. In other studies, quality is assessed within a frame of several months (during the same school year) of the outcome variables, that are themselves measured only once. These immediate outcomes are in contrast to *longitudinal* outcomes, an area that has been examined in a significantly smaller body of research. This paper refers to longitudinal outcomes as those variables that are measured in a later school year than when quality was assessed, or across the same year but controlling for initial differences in variables of interest. This area of research allows for predictability of outcomes from quality over time.

An additional distinction beyond immediate and longitudinal outcomes is the type of variables examined in relationship to quality. Most research on quality in child care settings concentrates on three types of variables: academic variables, social variables, and teacher-related correlates. Academic variables concern children's educational achievement, typically in literacy, language, or mathematics. Social variables refer to assessments of behavior, relationships, and communication; child-level variables outside of specific academic content that facilitate their interactions with other people and objects. Academic and social variables are often discussed as outcomes of quality, as opposed to teacher-related variables which are discussed as either predictors of quality or variables that tend to covary with quality. Teacher-related correlates generally concern more structural aspects of quality including teacher wage, teacher-child ratio, and classroom size. Research findings on the relationships between quality and academic and social variables both immediate and longitudinally, as well as with teacher-related correlates, are discussed in the following section.

Link to Child Outcomes

Immediate Academic Variables. Immediate academic variables have been investigated in studies using different instruments to measure quality, though the ECERS is one of the instruments most often used. For example, to examine the relationship between preschool quality and various academic skills, Bryant et al. (2003) used the ECERS as their assessment of classroom quality, thereby focusing on process variables, specifically those related to the environmental structure and materials. The researchers averaged ECERS ratings across several of the seven subscales, though they did not include the subscale related to adult needs, as an indication of global classroom quality.

To assess child abilities, researchers directly assessed children's competencies with standardized instruments examining language and literacy and numeracy abilities. Child data were collected two to four months after observers assessed the quality of the classrooms. Hierarchical modeling analyses, which accounted for the non-independence of children within same classrooms, revealed that, when controlling for gender, minority status, and poverty status, higher global classroom quality was significantly related to higher child receptive language skills, print awareness skills, book knowledge skills, applied math skills, and one-to-one counting skills. Their quality measure was not related to letter knowledge or story comprehension.

In an effort to examine how patterns of the relationship between quality and immediate academic variables look over a period of time as opposed to only within one year, as in the previous study, Burchinal et al. (2000) looked at within-time and across-time patterns of early childhood environment quality and children's academic skills. Both child assessments and quality ratings were collected at 12, 24, and 36 months of age. Researchers used the ECERS and the ITERS (Infant/Toddler Environment Rating Scale; Harms, Cryer, & Clifford, 1990), an instrument adapted from the ECERS but designed for use specifically with infant and toddler group care, to assess quality. Children were recruited for the study during their first year of life and assessed each year until they were three years of age. Global classroom quality was defined in this study as the total mean score of only the "child-related items" on the ITERS or ECERS. Instead of using the subscale scores to calculate the total average score on the ECERS/ITERS, the researchers only assessed those items directly pertaining to the children (excluding the items pertaining to adult needs) and averaged the ratings on those items alone. Each year this measure of quality was positively and significantly correlated with contemporaneous measures of children's cognitive development and receptive language skills, with the additional correlation between quality and expressive language found only at 24 months of age.

Researchers in this study (Burchinal et al., 2000) used a hierarchical modeling analysis to examine the pattern of the contemporaneous relationship between classroom quality and child variables at multiple ages. They concluded that, after adjusting for a host of family and child characteristics including gender, age, poverty status, and family environmental quality, in each year of the study, higher classroom quality was significantly and positively related to better academic variables in terms of cognitive development, receptive and expressive language, and communication skills, assessed within the same year as quality. Researchers reported a consistent pattern of the relationship between quality and academic variables across time periods in the study as children progressed through their first three years of life. Like previous studies, this work only related quality to child variables in the same year that they were measured, but this method was repeated over years as children aged to examine pattern consistency. Although researchers in this study had access to measures of quality and contemporan-eously-measured child-level variables at several different time points for the same children, researchers only examined the within-year relationship of these variables and did not examine the predictive nature of quality as it relates to outcomes in years following the initial quality measurement.

Immediate Social Variables. Also of interest to those who study the impact of child care is the relationship between preschool classroom quality and children's immediate social variables. Research on preschool quality and social correlates has tended to find mixed or no significant relationships but social competency is another concept that is measured in many different ways. In one of the few studies with positive results, Burchinal et al. (2000) identified the social variable through a direct assessment of children's social communication skills including gestural, vocal, and verbal skills, symbolic development, etc. Using the average ECERS rating from child-related items they found that quality in child care was consistently related to children's communication skills at 12 and 24 months of age. Higher quality during a single year was associated with higher scores during those years on measures of child communication.

However, Kontos et al. (2002) also examined relationships among child characteristics, classroom variables, and child interactions with objects and peers and did not find results similar to Burchinal et al. (2000). Their measure of quality was the average item score on the ECERS measured during a single two-hour visit. Kontos et al. reported that childcare classroom quality as measured by the ECERS was not significantly related to the total amount of children's observed complex interactions (defined by the authors as simple-social, complementary-reciprocal, cooperative-pretend, and complex-pretend interactions) with peers, or their creative or dramatic interactions with objects. Defining social variables as teacher ratings of child behavior, Bryant et al. (2003) also found no significant relationship between child care quality as measured by the average ECERS scores for child-related items and teacher ratings of children's social skills or problem behaviors after controlling for gender, ethnicity, and poverty status.

Wishard et al. (2003) defined social competency as children's relations with teachers and peers. In order to examine the effect of child care quality on this outcome, these researchers computed a quality construct through a principal component analysis, which combined scores on environmental quality (as measured by the ECERS), classroom emotional climate (as measured by the CIS and the Adult Engagement Scale), and teacher behaviors (from the AIS) to yield one quality factor. Along with this measure of quality, researchers included the child-teacher ethnicity pair status and teacher reported information on how their teaching reflected cultural practices, learning practices, socialization, and focus on peer relations into a regression model to predict child variables. The model did not significantly predict one of the social variables, the security of the teacher-child relationship. However, the model was significantly related to another outcome, children's competent behavior with peers. Though the overall model did predict children's unoccupied behavior, quality was not a significant contributor to that model.

The relationship between classroom quality and children's immediate academic and social variables cannot be definitively ascertained from the research reported in this section. Findings on academic correlates with quality are more consistent than reports of social correlates. However, researchers' measurement of quality remains an important issue. Although all of the studies reported here looking at the relationships between quality and immediate academic and social outcomes use the ECERS in their measurement of quality, only two of the studies use the instrument in the same way (Bryant et al., 2003; Burchinal et al., 2000). To complicate the matter, those studies do not use the ECERS in its entire form. The Kontos et al. (2002) study did appear to use the ECERS in its entirety, although the details of its use are not discussed in detail by the authors. The mixed results of research examining relationships between quality and child-level variables must be interpreted considering the measurement tools and methods of using those tools to provide a clearer picture of the findings as compared to each other.

Longitudinal Academic Outcomes. One powerful argument for the regulation of quality in early child care environments, assuming one could agree on a definition of quality, would be if variations in quality had lasting effects on important outcomes for children. Although research on immediate outcomes is important, longitudinal predictive analyses offer information on a more long-reaching effect of early experiences. In order to investigate the relationship between quality and development, Beller, Stahnke, Butz, Stahl, and Wessels (1996) conducted a study in Germany with public daycare centers serving children ages 6 to 24 months of age. Trained observers assessed the quality of the centers using the total ECERS mean, as well as individual subscale means, along with a time sampling instrument focused on caregiver behaviors during feeding, diapering/ toileting, and play activities. Because of the young age of the children in the sample, researchers modified some of the ECERS items to better assess children under three years of age, but no details were offered as to how this was done, how many items were altered, etc. Although observers rated classrooms using the entire ECERS, only six of the seven subscales were used for the analyses. The sixth subscale of the ECERS pertaining to children's social development was omitted due to its low Cronbach Alpha in relation to the other subscales. Developmental measures were assessed using a system based on a chart of developmental milestones. These milestones were scored through repeated child observations and time sampling techniques; the measures assessed variables such as autonomy, communication, affect, language and speech.

Quality measures were taken at the beginning of the school year, and child assessments were done both at the beginning and the end of the same year. Consequently, child post-test scores were adjusted for their pre-test scores in the data analyses, and all outcomes reported refer to the adjusted post-test scores. Although measures were taken during the same school year, the difference between this study and previously-cited studies examining immediate relationships between academics and quality involves the researchers' inclusion of pretest scores in their analyses. Adjusting for initial differences in academic proficiency allows for inferences about the facilitative nature of the quality measure on the outcomes. Researchers reported a high correlation between global quality from the ECERS and child's developmental status outcome ($r = .68$). In addition, ECERS total scores were broken apart into subscale scores and examined in terms of their relationships with child development status outcomes. Child academic developmental outcomes of language and cognition were significantly predicted by the ECERS subscale dealing with language and reasoning experiences but not by any of the other subscales.

An additional conclusion of the researchers was that ECERS scores were only predictive of children's behavior when that behavior was observed over time in the school environment. Neither the subscales nor the total ECERS scores were predictive of children's behavior when that behavior was measured through time sampling techniques. The researchers posit that the reason ECERS scores were predictive of continuously observed child behaviors but not of child behaviors measured with time sampling procedures may have been due to the different foci of the observation methods. Continuously observed child behaviors included a variety of situations across the observation period, much like the ECERS. In contrast, the time sampled behaviors focused on specific situations (feeding, toileting, and play). Although the finding that quality as measured by the ECERS was not predictive of child time-sampled behaviors might be related to the difference in situations observed, it could also be due to misuse of a measurement tool. With children under 2.5 years of age, the ITERS is generally used instead of the ECERS to assess center quality, as it is focused on the development of younger children. In addition, the exclusion of the social development subscale of the ECERS, which assesses items pertaining to a child's free play opportunities, might have led to the ECERS not predicting child behaviors as assessed by time sampling methods that focused on play and other specific routines.

The Beller et al. (1996) study employed the use of a pretest score to account for children's individual differences in the outcomes of interest at the start of the year. This technique allows for conclusions related to children's gain over the course of a year, improving upon other studies that only look at posttest scores (or one outcome score assessed during the same year as the quality assessment). The use of gain scores, or posttest scores relative to initial ability scores, makes a stronger argument for the effect of quality on outcomes than the use of posttest-only studies that cannot lean towards such causal claims. However, studies of relationships between child care quality and child outcomes necessitate the stance that the environment in which a child is located holds a certain degree of influence over the child's resulting behaviors. Assessing quality and child outcomes in the same year poses questions about how much of the observed child behavior was merely a state-dependent function of the environment they were observed in. Only studies that look at the prediction of quality to outcomes measured in later years, apart from the quality assessment year, can begin to hypothesize about the lasting impact of environmental quality.

Offering more information about the causal effects of quality was a study that also used the ECERS as a quality measure but was designed to assess the maintenance of the relationship between quality and child outcomes over time. Conducted by Sylva et al. (2006), this study of preschools in England examined the relationship between process quality and children's development as part of the Effective Provision of Pre-school Education study, a longitudinal study exploring the effects of pre-school provisions. In order to obtain a socially and ethnically diverse sample, programs were selected from five regions in England. One hundred forty-one centers were ultimately selected, representing the most common childcare options for children (parent cooperatives, private daycares, state nursery schools, etc.). Data were collected on 2857 children randomly selected from many preschools. Researchers used both the ECERS-R and the ECERS-Extension (ECERS-E; Sylva, Siraj-Blatchford, & Taggart, 2003) in three-year-old classrooms, and in all analyses, both the total average scores from each measure, as well as the scores on each individual subscale of each measure was considered.

The ECERS-E was developed as a supplement to the ECERS-R, recommended for use in combination with the ECERS-R. The instrument was designed with both English curriculum requirements and research on children's pre-academic skills in mind. Developers argued that the ECERS-R did not devote enough attention to the cognitive and pedagogical demands of the classroom necessary for children's intellectual and social development (Sylva et al., 2006). In addition, the ECERS-R was thought not to be a good enough measure of the cultural and academic diversity environment of the classroom. Three of the ECERS-E's four subscales each refer to a specific academic environment (literacy, science, and math), and the fourth subscale examines the diversity environment of the classroom; that is, how much and how well various cultures are represented in instruction.

On average, participating programs scored in the adequate to good range on the ECERS-R and in the adequate range of the ECERS-E when the mean total score across all centers on each measure was examined. Children were assessed for their cognitive skills and language knowledge at age 3 and again at age 5 using the British Ability scales. At age 5, children were also given a test of letter recognition and phonological awareness. Multilevel modeling was used in the analysis of the data. After controlling for age, pretest scores, and child and family background variables, quality as measured by the ECERS-E (both total and subscale scores) was significantly predictive of children's post-test scores on the pre-reading, general math concepts, and non-verbal reasoning skills assessed at age 5. Effect sizes, however, were fairly small, ranging from .108 to .166. No predictive relationships were found in terms of gains in spatial awareness and language. However, the total ECERS-R quality score was not significantly related to any of the child academic outcomes at age 5. One of the subscales, Interaction, was related to gains in children's general math concepts scores at age 5 (with an effect size of .199).

Other studies have examined the influence of preschool quality on specific academic outcomes in formal schooling using quality measures specific to academic content like the ECERS-E, rather than general environmental quality measures like the ECERS-R. Connor, Son, Hindman, & Morrison (2005) reported a relationship between preschool literacy environment quality and children's academic outcomes in first grade, although that relationship was not direct but instead exerted influence through preschool outcomes. Researchers assessed the quality of the preschool literacy environment by rating the classrooms on four observed teacher characteristics: questions asked by the teacher, choices offered by the teacher, teacher-facilitated learning, and teacher readings to the children. The authors gave no further information concerning the methods of observation in the preschool classrooms. In contrast to a measure like the ECERS that looks at global classroom quality, this method of assessing quality focuses specifically on the richness of the language/literacy environment of the classroom and the children's role in the learning process. Children's language skills, vocabulary skills, and letter-word recognition were measured in preschool, and their decoding skills, vocabulary skills, and phonological decoding abilities were measured in first grade. Through structural equation modeling, researchers found that the preschool environment measures they rated did not directly affect first grade academic outcomes. However, the preschool literacy environment directly affected children's preschool outcomes, which in turn predicted outcomes in first grade. Children who were exposed to higher quality preschool literacy environments had higher academic outcomes in preschool and subsequently had higher first grade academic outcomes. The continued effect of preschool quality on achievement scores beyond first grade was not examined.

Before conclusions can be drawn from the findings of this study, it is important to consider the authors' model for predicting child academic outcomes. If the indirect relationship between preschool quality and first grade outcomes is examined closer, it becomes clearer that a probable explanation for this relationship lies in the model. The authors report a direct relationship between preschool quality and preschool outcomes. As previously discussed in this paper, measuring quality and child-level variables in the same relative time period, especially without consideration of initial abilities, cannot lead to causal conclusions but may only represent a trend of higher-achieving children being enrolled in higher-quality preschool environments. In addition, the authors report a direct relationship between child preschool test scores and first grade scores. This finding isn't surprising considering that a child's scores at one time point and scores on the same measure at a later time point, given that the measure is reliable over time, *should* be highly correlated. If the researchers were looking for variables to explain the indirect relationship between preschool quality and first grade outcomes, they should have included another variable that was not as highly correlated with their outcome variable of choice.

Two studies, both pieces of larger longitudinal studies, have examined the effect of content-nonspecific childcare quality on academic outcomes well into elementary school. However, these studies have results that seem both conflicting and hard to resolve, one reporting a significant effect on math outcomes but not reading outcomes, and the other reporting the exact opposite. As part of a large-scale study of center-based child care and longitudinal child outcomes, the Cost, Quality, and Child Outcomes in Child Care Centers Study, researchers assessed quality and child outcomes in childcare, kindergarten, and second grade (Peisner-Feinberg et al., 2001). Preschool quality was measured with a combination of instruments, the ECERS, CIS, ECOF, and AIS, and the scores were combined through a principal components analysis to yield one composite index for each classroom. In kindergarten and second grade, modified versions of quality assessments were used. Researchers assessed children's receptive vocabulary, letter-word knowledge, and pre-math abilities. Hierarchical regression analyses revealed that, when maternal education, ethnicity, gender, age, quality in kindergarten and second grade, and teacher-child relationship measures were included in the model, only math outcomes were significantly predicted by preschool quality. Researchers did not include children's prior test scores from preschool through first grade as covariates in the model. Children from childcare environments of higher quality tended to have higher math outcomes in second grade. Researchers did not find a predictive relationship between childcare quality and vocabulary or letter-word outcomes.

In contrast to the findings of Peisner-Feinberg et al. (2001), a similar study examining the effects of child care experiences through sixth grade did not find a relationship between quality and math outcomes (Belsky et al., 2007). This study was part of the National Institute of Child Health and Human Development Study of Early Child Care and Youth Development that began in the 1990’s and examined the effect of child care on longitudinal outcomes, following children from birth through sixth grade. Researchers observed the quality of childcare that children experienced at 6, 15, 24, 36, and 54 months of age using the ORCE, a measure of the quality of the caregiver-target child interactions. During the formal elementary years, researchers used the COS to measure quality in first through fifth grade classrooms that included target children. Both the ORCE and the COS use rating scales to assess the quality of the environment based on observation periods of 44 minutes (two cycles in child care and first grade classrooms; eight cycles in third and fifth grade classrooms), and both examine the individual target children and the environment around those children. Children’s outcomes were assessed in preschool, first, third, and fifth grade, mainly concentrating on either a letter-word knowledge task or general reading task (based on age), an applied problems mathematical skills assessment, and an assessment of expressive vocabulary knowledge. Controlling for child and family demographic measures, the only fifth-grade outcome that was significantly predicted by child care quality was expressive vocabulary. Children who experienced higher quality in child care had slightly higher expressive vocabulary scores in fifth-grade. This relationship was not found for the reading measure or the math measure, although reading scores were predicted by quality through kindergarten but not beyond.

Upon first glance, it can seem confusing that two large longitudinal studies of quality and academic outcomes reported such contradictory findings. Indeed, without further analysis of the methods used in those studies, it would be difficult to determine the implications of child care quality on specific longitudinal outcomes, especially considering that each study used different instruments to assess quality, followed children for different amounts of time, conducted different data analyses, and so forth. However, one shared characteristic stands out; neither study was experimental. Rather, both studies have tracked children through their early years, measuring characteristics of their home and academic experiences as they transitioned to and progressed through school. When participants can be randomly assigned to conditions and the characteristics of each group assessed to ensure group equality, one can be much more confident in the findings of the research. When random assignment and group design are not utilized, differences in study groups, both before and after attrition, and differences in the experiences of those participants in study classrooms can vary widely. Though each study, especially the study by Belsky et al. (2007), attempted to measure and control for factors associated with demographics and family life to minimize differences in outcomes due to initial differences, neither study employed experimental techniques allowing for causal inferences. For this reason, the findings of studies not utilizing these design strategies must be carefully scrutinized for any reliable implications for practice.

Longitudinal Social Outcomes. In addition to examining the relationship between quality and long-term academic outcomes, researchers are also interested in what later child social behaviors are associated with early educational experiences. While no one study seems to link any particular measure of early educational quality to all the social behaviors assessed in that study, some research has found a link between some measures of quality and some longitudinal social outcomes. Researchers in England found that, while the ECERS-R total score or subscale scores did not have the predictive power related to academic outcomes that the ECERS-E total score or subscale scores did, preschool quality, as measured by the ECERS-R total and subscale scores separately, predicted some social outcomes in first grade (Sylva et al., 2006). The total ECERS-R score from preschool observations explained a significant amount of variance in the first grade teachers’ ratings of cooperation and conformity. In addition, the Interaction subscale from the ECERS-R was associated with teacher-rated independence and concentration as well as peer sociability in first grade. ECERS-R scores indicating higher quality in preschool were related to better teacher-rated pro-social outcomes for children in first grade. On the other hand, neither the total quality score nor any subscale score predicted teacher ratings of child anti-social/worried behavior.

Often researchers using the ECERS as a quality measure have examined the utility of both the total score as well as subscale scores to explain variance in outcomes, as in the Sylva et al. (2006) study. However, where Sylva et al. found the total ECERS-R score to be predictive of some social outcomes, other researchers have only found a relationship between subscale scores and outcomes. Beller et al. (1996) found that neither the total infant/toddler ECERS scores nor any of the subscale scores predicted child social behavior one year later using time sampling methods to assess child skills in classrooms. However, two of the ECERS subscales significantly predicted children's social-emotional ratings and level of play observed when child behavior was assessed through continuous participant observation.

Another study found a similar relationship between preschool quality and second grade teacher-ratings of child problem behaviors (Peisner-Feinberg et al., 2001). As in the Sylva et al. study (2006), the ECERS total score was used as a measure of preschool quality, but it was used in conjunction with several other measures, and scores were eventually combined across measures to form one composite score. In this study, the quality composite score was not found to be predictive of any type of social behaviors. One measure that could be thought of as quality used by these same researchers was preschool teachers' ratings of teacher-child closeness; these preschool ratings were related to second grade teachers' ratings of children's problem behaviors. Other social behaviors examined, cognitive/attention skills and sociability, were not significantly predicted by either measure of quality. One wonders if the preschool teacher ratings were more a measure of early child problems than a measure of environmental quality.

While the above-mentioned studies did find some predictive relationships between quality in the various ways it was measured and child social outcomes over time, a 2007 study by Belsky et al. did not find any relationship at all. Researchers in this study measured quality in child care using the ORCE. Hierarchical modeling did not reveal any significant predictive relationship from child care quality to teacher-reported child externalizing behavior, social skills, conflict, social-emotional skills, or work habits in sixth grade. Belsky et al. reported that, in contrast to other studies' findings of relationships between preschool quality and later child social outcomes, they did not find child care quality to be associated with children's social behaviors in later grades. They found child care to be associated with higher teacher-ratings of externalizing problems and teacher-child conflict through fifth grade. Interestingly, however, the relationship was between the *quantity* of time children had spent in child care not the *quality* of the environment. As mentioned previously, in the same study, Belsky et al. reported that child care quality was a significant predictor of children's vocabulary scores in fifth grade. According to their study, *quality* of child care was predictive of academic outcomes but *quantity* of child care was predictive of social outcomes in formal school years.

The authors posited that this relationship between quantity and social outcomes might be a result of the insufficient training of the child care teachers, ill-preparing them to deal with child behavior problems, combined with a lack of time to deal with such problems due to their focus on the academic environment of the classroom. However, if teacher training and/or time were truly behind the effect on child outcomes, one would think that under-prepared teachers would also affect the *quality* ratings of the centers, and in turn quality would help predict child social outcomes, as well. An important issue to consider when making such hypotheses is that quality was measured at the classroom level through a mean score on the ORCE instrument. In contrast, quantity of care was assessed on an individual student level through parent report. Too little is known about how many times a classroom would have to be observed to obtain a stable measure of something like quality. Quantity can be measured fairly precisely if parents are contacted regularly.

Teacher Characteristics Predicting Quality

A large amount of research has been conducted that examines the relationship between the quality of early educational settings and teacher-related or structural (regulatable) characteristics such as teacher pay, center fees, and teacher training. In this case, the interest is in determining whether these factors account for variation in quality. This research has not yielded consistent results. Many studies report positive correlations between quality and regulatable aspects of early education across infant/toddler care, preschool, and kindergarten, while others report null or negative relationships among the same variables. In addition, sometimes researchers do not report the significance levels of the relationships found, leading to difficulties in comparing results across studies.

Some studies have reported that teacher wage is the strongest correlate of classroom quality (Phillips et al., 2000; Scarr et al., 1994; Whitebook, Howes, & Phillips, 1990; Phillipsen, Burchinal, Howes, & Cryer, 1997). Phillips et al. measured quality with a combination of scores from the ITERS/ECERS (depending on the age of the children in the classroom), and the Assessment Profile. A maximum likelihood factor analysis of the scores revealed that one factor explained 40.3% of the variance, and so items from all three instruments were standardized and averaged together for one total score per classroom. The researchers found that both infant/toddler classroom quality and preschool classroom quality were more strongly associated with teacher wage than any other measured structural variable (ratio compliance, group size, actual teacher-child ratio, teacher education, teacher training, and parent fees).

Scarr et al. also reported that teacher wage had the strongest relationship with childcare quality as measured by total scores on the ITERS/ECERS (*r*=.59) and the Assessment Profile (*r*=.51). Teacher wage was correlated with quality measures more strongly than ratio, group size, teacher training, teacher education, and staff turnover. In contrast, Pianta et al. (2005) reported no significant contribution of teacher wage to the prediction of quality in preschool classrooms, regardless of whether wage was entered as either the first or last block in a regression model with other variables (including state, ratio, teacher training, teacher experience with preschoolers, etc.). Researchers in this study measured quality with the ECERS-R and the CLASS, but rather than analyzing those quality scores as total instrument scores, the items on each instrument were divided into two separate factors for each measure, based on factor analysis. The two factors for the ECERS-R were labeled as Teaching and Interactions (including items related to teacher-child interactions, language development, discipline, etc.) and Provisions for Learning (including items focusing on the materials provided and the physical space). For the CLASS, the two factors were labeled as Emotional Climate and Instructional Climate.

Similar to Scarr et al., the researchers in the Pianta et al. study did find a small but significant correlation between one of the ECERS-R factors (Provisions for Learning) and teacher wage (*r=-.20),* but this relationship was negative; higher wage was associated with lower ECERS-R Provisions scores. However, when the variables were entered into regression analyses to predict global quality from a various teacher variables and controlling for center and class characteristics, teacher wages were not a significant predictor of quality as measured by either the CLASS or ECERS-R factor scores.

Another block of teacher-related variables often examined in combination with quality includes teacher education, teacher experience, and teacher training. While some or all of these variables have been reported to be positively correlated with quality at the infant/toddler level (Phillips et al., 2000; Phillipsen et al., 1997) and preschool level (Phillips et al., 2000, Buysse, Wesley, Bryant, & Gardner, 1999; Cassidy, Hestenes, Hegde, Hestenes, & Mims, 2005; Whitebook et al., 1990; Phillipsen et al., 1997), other researchers have found no relationships between teacher education, training, and experience and measures of quality (Phillips et al.; Scarr et al., 1994; Pianta, LaParo, Payne, Cox, & Bradley, 2002). Some of these studies assessed quality in the same way (Buysse et al. and Whitebook et al. used the total ECERS scores), while others examined quality using different measures or combinations of measures (Phillips et al. used a combination of three measures, and Pianta et al. used the COS-K).

Recently, using seven of the largest studies of early childcare and education, Early et al. (2007) conducted a secondary data analysis to examine the effect sizes obtained by those studies relating teacher education to quality and child variables. Within the seven studies, six studies defined quality by the total mean ECERS-R scores. The seventh study measured quality with the ORCE. Early et al. concluded that the evidence from the combined studies did not offer support for the belief that teacher education (neither years of school nor major in school) was related to classroom quality or children’s academic progress in receptive language or prereading skills (measured identically in six of the seven studies), or early math skills (measured the same in all seven studies).

Conflicting information about whether or not certain teacher characteristics are related to quality poses a dilemma for policy makers because these characteristics are among the easiest to regulate. This problem is exacerbated by the fact that different results can be found when different measures are used. For example, Pianta et al. (2005) found that pre-kindergarten teacher education and training did significantly predict quality as measured by the author-defined factors resulting from a factor analysis on both the CLASS and ECERS-R. The Emotional factor of the CLASS and Provisions factor on the ECERS-R were predicted by the teacher education variable. Neither the Instructional factor from the CLASS nor the Interactions factor from the ECERS-R were related to teacher education and training. The logic model being tested is that teacher characteristics such as educational degree obtained and amount of training in early childhood are related to classroom quality, and classroom quality in turn is related to child variables. But the research suggests that different measures of quality yield different results, and sometimes even the same measures produce different findings.

Other structural quality variables of interest to researchers have been group size and teacher-child ratio. Some studies have found significant relationships between quality and ratios; higher teacher-child ratios were associated with higher quality measured with a combination of ECERS, ITERS, and Assessment Profile scores (Phillips et al., 2000) or with the ITERS, ECERS, and CIS analyzed separately (Phillipsen et al., 1997). In addition, some studies report small to moderate correlations between these variables but do not report significance levels; for example, Scarr et al. (1994), using two measures of classroom quality in 363 classrooms, reported a correlation between ratios and quality measured by the ITERS/ECERS as .36 and a correlation between ratios and the Assessment Profile scores of .31. In contrast, Cassidy et al. (2005) found a nonsignificant (but in the negative direction) correlation of -.12 between total scores on the ECERS-R and teacher-child ratio. In terms of the relationship between various measures of quality and group size, results have also been contradictory, but have most often revealed no significant relationship. Scarr et al. reported a correlation of -.10 between these two variables. Cassidy et al. reported a similarly small correlation of .05. Phillips et al. did find group size to be a significant predictor of classroom quality measured with an overall quality score taken from ratings on the ITERS, ECERS, and Assessment Profile, but only in classrooms serving toddlers and not in infant or preschool classrooms.

Quality in Older Grades

Although kindergarten is arguably different from preschool in terms of focus and can be thought of as part of formal schooling years as opposed to early childhood education, research involving the relationships between classroom quality and other variables can be helpful in thinking about this complicated issue. To examine the relationship between global classroom quality and kindergartners' immediate academic progress, researchers have often used teacher-ratings of child competencies as dependent variables in their analyses (Pianta et al., 2002; Stipek & Byler, 2004). Though quality was measured in different ways, both Pianta et al. and Stipek and Byler conducted studies attempting to link classroom quality to teacher-ratings of children's academic achievement. Pianta et al. measured quality using the Classroom Observation System for Kindergarten (COS-K; NCEDL, 1997). The COS-K uses a combination of time sampling and global ratings to assess both the classroom as a whole and individual children's experiences within that classroom through direct observation. Stipek and Byler used the Early Childhood Classroom Observation Measure (ECCOM; Stipek, 1996) to assess classroom quality, a measure originally developed for the purposes of this study. The ECCOM focuses on two aspects of instruction: constructivist approaches and didactic approaches. Pianta et al. reported that kindergarten quality, included in a regression model including maternal education and family income, explained 15% of the variance in teacher-rated literacy skills and 17% of the variance in teacher-rated math skills. Stipek and Byler concluded that teachers' more constructivist practices were unrelated to teacher ratings of children's academic competence and teacher ratings of students' self-directed learning. However, didactic teaching practices were related to teacher ratings of students' math skills and self-directed learning. Negative correlations were found for each of those relationships; a teacher who demonstrated more didactic practices was more likely to rate his/her students lower in math achievement and self-directed learning strategies.

Both the Pianta et al. (2002) study and the Stipek and Byler (2004) study had similar issues that make it difficult to infer true meaning from their findings. Both studies sampled a very small number of children from each study classroom. Pianta et al. included only one target child from each classroom, and Stipek and Byler reported an average of 1.84 target children in each of their study classrooms. Although this allowed for data analyses to rightfully assume independence of children across classrooms, it did not necessarily provide an accurate picture of the experience of the majority of children in the classroom who were not targeted. In addition, neither study gathered data pertaining to initial child academic achievement. These studies illustrate the difficulties of using a posttest only design. One cannot be sure if the relationships are illustrative of individual differences among children or connected to classroom quality. A last issue shared by both studies is that teacher-ratings were used for child competency measures. Rather than implying that higher quality classrooms lead to higher child academic achievement, one might conclude from these studies that in classrooms with higher observed ratings, teachers also liked their children more. A more in-depth view of a child's experience of quality in a classroom might be the combination of environment or teacher quality and teacher views of student abilities, and that cluster could then be used to predict other independent measures.

Research findings involving immediate academic outcomes and their relationships to preschool quality are somewhat similar to the findings regarding quality in higher grades. However, in contrast to the null findings relating preschool quality to immediate social outcomes, research on kindergarten quality may reveal a slightly different view. Stipek and Byler (2004) reported that, although the constructivist practices aspect of their kindergarten quality measure (from the ECCOM) was unrelated to children's social outcomes, didactic practices were negatively related to teacher ratings of their closeness to the target children. More constructivist practices were positively related to ratings of teacher-child closeness, but not significantly so. However, due to the small number of children sampled from each classroom (average of 1.84), correlations might be the result of response bias; more didactic teachers were more likely to rate their children lower. Without more children in each classroom, there is no way to check for this possibility.

Two studies relating kindergarten quality to children's immediate social variables using the same measure of quality were conducted by Rimm-Kaufman, LaParo, Downer, and Pianta (2005) and Pianta et al. (2002). Both studies used the COS-K to measure classroom quality but found conflicting results. Rimm-Kaufman et al. reported that quality accounted for a small but significant percentage of variance, between 2 and 6%, in child social behaviors of cooperation with peers and occurrences of off-task behavior during whole-class activities. Researchers did not find quality to be a significant predictor of child aggression toward peers, compliance, or occurrences of on-task behavior in any setting. In contrast to the null findings regarding children's on-task behavior, Pianta et al. reported that kindergarten quality was a significant predictor of children's on-task behavior and teacher-reported social competence when quality was entered into the regression model along with maternal education and family income. Contradictory relationships like this often emerge in research evaluating child care quality. In order to minimize the confusion, it is necessary to acknowledge the difficulties that arise in assessing quality and think about how to deal with these difficulties in future research.

Studies in quality involving kindergarten and older grades can provide some insight into relationships that might exist between quality and child variables at younger ages. However, poorly-designed studies that do little to account for the difficulties coinciding with evaluating quality and its correlates offer limited implications for practice with children at any age. Rather, these studies call for further research on quality using study designs that incorporate strategies to deal with classroom evaluation difficulties.

Difficulties in Determining the Importance of Quality

Several design issues arise in studies attempting to evaluate environmental quality and its relationships with other variables. The primary issue concerns the inference of causality often found in short term studies. Studies that assess child skills at exactly the same time (or within the same year) that quality is measured often report significant correlations between the two variables and use those correlations to speak to the causality of quality. In fact, the researchers often refer to the child measure as "outcomes," thereby implying that the environmental measures are the predictors. Though these correlations display potentially interesting information, no child-level or classroom-level variables are typically controlled for in the analyses. When children's entering academic status in a study year is not assessed, one cannot conclude that children who were exposed to higher quality each year *developed* better academic competencies. An alternative but also plausible solution based on correlations is that children tended to receive the quality of care related to their entering abilities. Parents of higher achieving children were likely better educated, employed in higher earning jobs, and able to obtain more costly (and higher quality) child care. This problem persists in research that does not consider children's gains as outcomes but only looks at post-intervention scores or, alternatively, does not include children's pre-intervention scores in the analyses.

Another area of concern when interpreting the results from studies of quality is the method of analysis that the researchers used. When the interest is to examine the effect of classroom variables on child outcomes, it is better to sample more than one or two children from each class. Using only a small number of children in a classroom tends not to give an accurate representation of the true variability in child outcomes from the entire classroom, but rather the data are representative of a small and perhaps unrepresentative portion of the class. However, when several children from each classroom are included in a study, simple correlations or regressions do not fit the data either. These analysis techniques assume independence of study units. That is, a regression would automatically assume that children in the same class were independent from one another, as independent as a student from one classroom would be from a student in another classroom. This logic is faulty; despite variation among them, children from the same classroom might very well be more similar to each other than they are to children from another classroom because of their exposure to the same set of experiences within their class. Therefore, it is recommended that researchers employ a hierarchical modeling technique in their data analyses to account for the nesting of children within classrooms (Van Horn, Karlin, Ramey, Aldridge, & Snyder, 2005).

A third difficulty in interpreting the results of studies of quality concerns the measurement tools employed. Causality inferences and analysis issues, while certainly areas of concern when interpreting study results, both involve the added complication of the measurement tool used. Even if all other study concerns are alleviated, issues surrounding the instrument that is used to yield the study's main dependent variable can permeate the other characteristics of the research. A study that focuses on quality and its links to other variables must be able to place confidence in the measurement tool used. The same truth holds when attempting to characterize the relationship of quality and other measures across studies.

Results from studies attempting to find links between early educational quality and child outcomes clearly conflict. However, it is difficult to compare the results of such studies, chiefly because of their use of different instruments to assess the quality of the environments. As discussed earlier, instruments all purporting to assess quality in early childhood classrooms employ different methods, have different psychometric properties, and stem from different theoretical backgrounds. Rarely do practitioners or researchers agree on what constitutes quality in an early childhood education setting, evidenced by the groups of research articles claiming to have found relationships between classroom quality and other variables but, upon closer examination, have measured quality in completely different ways (e.g. Kontos, Burchinal, Howes, Wisseh, & Galinsky, 2002; Wishard et al., 2003). A relationship found between quality as measured by one particular instrument and an outcome might very well be a reflection of the content of the instrument, and that relationship might not be found when using a measure of quality that focuses on different aspects of the classroom, even if both measures employ direct classroom observation.

Most people would argue that the quality of early childhood school environments s*hould* matter. Children who are exposed to educational environments that are high quality should have a better chance of succeeding in formal schooling than those children whose experiences are poor in quality. However, as the research shows, this relationship is often not found. It is possible that one of the reasons for this lack of prediction from quality assessments is because of the differences in instrument construction, instrument use, and/or theory behind the measures. In the following section, several instruments designed to assess quality that were used in the studies mentioned earlier are compared and contrasted to illustrate differences that could lead to the conflicting results from those studies.

Examples of Quality Measures in Detail

Three of the often-used assessments of quality in early childhood educational environments, the CLASS, ECCOM, and ECERS-R, stem from different theoretical backgrounds and vary in their instrument design. All of them use direct observations in classrooms through the use of trained observers. The CLASS has a different approach for measuring quality than the other two. The CLASS does not assess the physical or structural features of a classroom, but rather looks at the nature of the interactions between teachers and children and children and their peers. The development of the instrument stems from developmental theory, stressing teacher-child interactions as the principal system operating to allow for children's learning and development. The CLASS is comprised of four broad domains (Emotional Support, Classroom Organization, Instructional Support, and Student Engagement), within which 11 items are scored. Observers are supposed to remain in the classroom for a three-hour period, during which 20-minute cycles are observed and then coded for 10 minutes (however, the developers recommend a minimum of only four cycles, indicating observers could stay for as few as two hours; Pianta, LaParo, & Hamre, 2006). Items are scored on a scale from 1-7, broken down into low-range, mid-range, and high-range. Ratings for each scale are based on a set of descriptors that observers should be attuned to during their observation cycles. The developers of the CLASS emphasize specifically and emphatically that the descriptors are not to be rated as a checklist. Rather, observers are to use the descriptions to form a global impression.

The ECCOM also uses direct observation for assessment, but was designed and developed differently from the CLASS. The theory behind the ECCOM draws from both constructivist theory and traditional learning theory (Stipek & Byler, 2004). Developers of the ECCOM borrowed from the constructivist view that children are active learners and constructors of knowledge, as well as from the view that direct instruction in combination with practice are keys for learning and development. Because of these two views, the ECCOM assesses the degree to which teachers' practices reflect both constructivist and didactic methods. Observations focus on the methods of the instructor rather than the content delivered. In terms of its composition, the ECCOM scores are based on 32 items (17 constructivist and 15 didactic) that are rated on a 1-5 scale ("rarely seen" to "predominates"), and the items are grouped into three subscales: Climate, Management, and Instruction. The score for each item is based on an estimate of the percentage of time that those practices were observed. Like the CLASS, the recommended observation period is a minimum of three hours.

A third measure of quality, the ECERS-R, is an assessment of the physical environment of the classroom, and to a lesser degree, the warmth of interactions between the teacher and child, and was designed to examine the global process quality of the environment. This instrument is composed of 43 items within seven subscales. Each item is rated during live observation on a scale from 1-7, or Inadequate to Excellent. Beneath each of the odd numbered points is a list of behavioral indicators. Unlike the CLASS, the scoring of each item is based completely on the number of individual indicators checked as "yes" or "no" corresponding to each item. If an observer scores more than half of the indicators under a rating number as "no," that item cannot be scored any higher than the previous rating number. The ECERS-R is designed for use in classrooms serving children 2.5 through 5 years of age and requires an observation period of at least two and one half hours. In addition, several items on the scale might necessitate a brief teacher interview at the conclusion of the observation in order to determine the appropriate response. Researchers wanting to assess the global process quality of a classroom will often use the ECERS-R without the seventh subscale, Parents and Staff, because those characteristics included under that heading are usually thought to be aspects of structural quality rather than process quality (Cryer, 1999). The next section of this paper will delve deeper into the design and use of this instrument.

The ECERS: An Example of a Classroom Quality Measure

As previously stated, the ECERS is the most widely-used measure to assess the quality of early childhood environments. According to the developers' website, the ECERS-R and its sister measures looking at different populations (ITERS; Family Day Care Rating Scale, FDCRS; School Age Care-Environment Rating Scale, SACERS) have been or are currently being used in 13 research and evaluation projects, both nation-wide and internationally, 19 government regulation projects, nine teacher training projects, and 65 program improvement projects (*Environment Rating Scales*, n.d.). Although these numbers seem large, they are only a sample of the projects using the rating scales. The information provided on the website refers to those projects that have been brought to the attention of the developers. In truth, one would imagine that the use of the rating scales is far more extensive that what is represented here. The ECERS-R is most often cited in research as a measure of global process quality, examining aspects of the environment pertaining to what children are exposed to, including materials, health and safety routine practices, and general interaction aspects.

In the past ten to fifteen years many states have begun to use the ECERS as either a required or optional piece of their childcare licensing process. By linking quality measures such as the ECERS to state licensing and rating protocols, a center's score on the instrument becomes closely tied to the status and funding of the program. In many states, higher ratings on the ECERS mean higher rates of reimbursement for child care centers. The first portion of this paper focused on the existing research attempting to link quality to child- and teacher-level variables. Many of the studies reported used the ECERS to assess program quality. Despite the contradictory findings of research on quality and its correlates using the ECERS, the instrument is currently widely used in ways that directly affect the status and funding of child care centers. The following section highlights two examples of states that use the ECERS in their licensing practices.

The first example is the state of Tennessee. In Tennessee, all child care providers seeking licensure initially and again every year after licensure is obtained are required to participate in the Child Care Evaluation and Report Card Program. This program is designed to provide detailed information to parents about the quality of childcare their child is receiving (*Safe, Smart, and Happy Kids, 2005*). The licensing process for child care centers is based on seven domains: director qualifications, professional development, compliance history, parent/family involvement, ratio and group size, staff compensation, and program assessment.

For the program assessment component, state evaluators use the ECERS-R (or the age-appropriate version for infant and toddler programs, the ITERS) to evaluate the environments of childcare programs. There is no minimum standard score on the rating scales to qualify for licensure, but the scores are generally considered in combination with scores in the other six domains to determine a center's licensure status.

In addition, all licensed child care centers are given the option of participating in the Star-Quality Child Care Program, a program designed to acknowledge settings that go beyond the minimal licensing standards.   In order to be eligible for the Star Quality program, a center must have obtained an overall average score (across all seven subscales of the ECERS-R or ITERS) of a 4.0 or higher. More stars (on a scale of 1-3) are indicative of higher quality programs. A center ultimately receives their overall star rating based on the rating scale's associated stars, as well as stars awarded in the other six domains. More stars are also equated with higher reimbursement rates per child that participates in the Child Care certificate program. A one-star rating results in a 5% increase above the base reimbursement rate for each participating certificate child. Two stars are associated with a 15% increase above the base, and three stars result in a 20% increase.

Ohio is the second example of a state using the ECERS-R to make decisions about quality and therefore reimbursement. Ohio incorporates the rating scales scores into their voluntary program of quality assessment. The rating scales, however, are not used to determine licensure qualification. Licensed child care facilities in Ohio are invited to participate in the state's Step Up to Quality program (*Step Up to Quality,* n.d.). The program incorporates five areas of requirements: ratio/group size, staff education and qualifications, specialized training, administrative practices, and early learning (the area involving the rating scale assessment). Like

Tennessee, Ohio uses a star rating system to indicate the quality of a participating center as part of the early learning domain. Although the use of rating scales is not required to receive a one star rating, centers must perform an annual self-assessment with the appropriate scale (ITERS, ECERS-R, ELLCO, and/or the School-Age Care Environment Rating Scale; SACERS) to receive two stars. A center must decide which of the scales to use based on the ages of the children served, and must use all appropriate scales for that age group. The ELLCO (Early Language and Literacy Classroom Observation; Smith & Dickinson, 2002) is designed for use with children from age three to third grade, so if a center serves preschool-age children, both the ECERS-R and the ELLCO must be completed. If a center serves children birth through preschool, the ITERS would need to be completed in the infant and toddler rooms, and the ECERS-R and ELLCO would be used in the older classrooms. Three stars require not only the annual self-assessment but also evidence of an action plan for changes resulting from the rating scale assessment. The higher star rating a center receives (based on the five program domains) is tied to distributed Quality Achievement Awards. More stars can lead to higher monetary awards based on the percent of subsidized enrollment of each center.

Tennessee and Ohio are just two examples of states that use the ECERS to determine some of the funding child care facilities can receive. A center's scores on the rating scales can mean better materials for the children served, as well as higher pay for the staff. Even centers that do not enroll many state-subsidized children (whose reimbursement rates go up with each star) are invested in achieving the most number of stars as a recruitment tool for families. When a particular instrument is so closely linked to the lives of citizens and to the way a community cares for its youngest ones, it is critical that the instrument be carefully evaluated so that everyone affected by its use can have confidence in the scores it yields. The next section looks specifically at the ECESR-R and the measurement issues associated with its use in classrooms.

ECERS Composition

An issue concerning the design of the ECERS involves the nature of the scoring system. The ECERS-R is comprised of over 400 individual indicators, organized into 43 items under seven subscales: Space and Furnishings, Personal Care, Language-Reasoning, Activities, Interaction, Program Structure, and Parents and Staff. There are multiple ways to score the instrument, and although the developers mention alternate options in the instrument's instructions, one method is nearly always used, which involves a shorter observation period and fewer indicators to be scored. The traditional scoring method is described later in more detail. However, the consequences of scoring a classroom in the alternate ways are unclear. If multiple ways to score an instrument are possible, it is of import that those different methods offer comparable results.

The ECERS is a criterion-referenced measurement scale (Bagnato, 2002); however, there is little evidence that it operates as one. According to Horne (1984), a criterion-referenced scale is designed to compare one entity's scores to a set standard score, rather than compare one entity's scores to another entity. Although the ECERS does reveal a score that is indicative of a center's quality as compared to the standard of quality held by the instrument's developers, ECERS scores are frequently used to compare the quality from one program to another. With the use of the ECERS in current policy, scores are often reported to parents so that they can decide which center they would choose for their child to attend. This comparison among programs is not aligned to the theory behind criterion-referenced measures.

Another issue surrounding the scoring of the ECERS-R concerns the individual behavioral indicators. With the way that the ECERS is scored, if a classroom receives a one on an item (representing the lowest possible score at the Inadequate level), that classroom still might have some of the indicator items beyond the lowest anchor present. However, if enough of those indicators are absent, a score of one still may be obtained. One classroom may receive a score of a three but have many indicators pertaining to a score of five checked, while another classroom that also has a score of three on an item might not have any of those higher indicators. This fallible classroom comparison leads to another issue with the ECERS scoring. An instrument comprised of over 400 individual indicators can offer a very detailed picture of the quality components of a classroom. However, when those indicator scores are reduced to a total average score on a seven-point scale, a vast amount of variability in quality as evidenced by individual indicators is erased. Two classrooms that have identical ECERS scores may in actuality be very different from one another when the individual indicators present in each of those classrooms are considered.

ECERS-R Items

Regardless of the scoring method used, another critical issue for an instrument so steeped in the policy realm is whether the tool is examining aspects of a classroom that are thought to be important to quality. The developers of the ECERS-R leave their theoretical stance out of the instrument. It is therefore up to the user to ascertain the definition of quality that the developers hold based on the items that are included in the scale. However the instrument is scored, the items that it looks at must be accepted by various communities, namely the research, education, and policy communities, so that the results of the assessment yield information about the quality of an environment that can be valued by the users. A large question concerning the use of the ECERS-R, then, is whether the instrument is even assessing classroom characteristics that are widely-thought to be components of quality.

ECERS-R Psychometrics

Another major issue with such a popular instrument, especially, is whether it has been shown to be a reliable and valid measure. Like the majority of rating scales assessing quality, however, the ECERS has had limited psychometric evaluation (Perlman, Zellman, & Le, 2004), and even the evaluation that exists has been conducted primarily for the traditional scoring method.

The original ECERS was made up of only 37 items but still grouped into seven subscales. The validity of the instrument was examined only through expert opinion, and the grouping of items into subscales on the original instrument was done based solely on face validity. Since the development of the ECERS, some researchers have questioned the grouping of items into seven discrete subscales. Some research indicates that instead of measuring seven distinct characteristics of quality as the subscales might indicate, the ECERS-R actually examines one global quality factor, and that much of the information assessed in the instrument's indicators is repetitive (Scarr et al., 1994; Perlman et al., 2004). In a 1994 study of 363 early childhood classrooms in three states serving children from one to five years of age, Scarr et al. conducted a factor analysis of the ECERS at both the subscale and the individual item level. Their analyses revealed one global factor on which all items loaded, with an eigenvalue of 4.82 that explained 69% of the common variance. In addition, given the high intercorrelations among items, the researchers conducted further analyses. They randomly selected three sets of twelve items from the ECERS to determine whether any one of those smaller sets might be used instead of the full instrument and obtain similar results. Correlations between the entire ECERS and each of the three randomly selected subsets of items were high, ranging from .93 to .95.

In an attempt to replicate the findings of Scarr et al. (1994), Perlman et al. (2004) collected quality data on 326 classrooms using the ECERS-R. These investigators found similar results except their factor analysis revealed not seven discrete factors but three. In addition, the first factor had an eigenvalue of 13.85, which was seven times larger than the eigenvalue of the second factor, and explained 71% of the common variance. The authors also wanted to explore the finding by Scarr et al. that any random set of 12 items from the ECERS was comparable to the entire instrument. Perlman et al. went a step farther and purposively selected subsets of items; one subset was comprised of 24 items selected for their ease of measurement, generally pertaining to aspects of the physical environment. The second subset was comprised of 10 items selected by a group of childcare practitioners. These items were more focused on interactions. In both cases, the subsets proved to be highly correlated with scores on the entire ECERS-R (.92 for the subset of 24 items and .88 for the subset of 10 items). Correlations remained high even when items common to both the subset and the total instrument were removed from the analysis.

Although the makeup of the ECERS is important for determining whether one or multiple dimensions of the quality of an environment are being assessed, as well as determining whether a much shorter instrument requiring less training and money would yield comparable results, issues of instrument reliability and validity are of equal or greater importance. Although the general terms of scale reliability and validity encompass many different characteristics of the measurement, this section focuses specifically on the temporal stability and predictive validity of the ECERS.

Temporal stability involves reliability over time; an instrument that is not reliable from one time period to the next within the same classroom has serious implications for its use. Temporal stability is most commonly assessed with the test-retest method, involving the administration of a test at one time period correlated with the same test administered to the same group of participants after a certain amount of time (Crano & Brewer, 2002). A study of 21 classrooms that were each assessed in the winter and in the spring of the same year with the ECERS-R revealed no significant correlation between a classroom's scores for each time point (Farran, Lipsey, Hurley, & Bilbrey, 2006). Thus these researchers found that the ECERS-R did not reliably assess the quality of a classroom twice in a four-month period. The test-rest reliability of the ECERS is rarely examined or reported. In fact, the only other reference this author found to the temporal stability of the ECERS appeared in a review of six studies of state pre-kindergarten programs conducted by the National Center for Early Development and Learning, or NCEDL (Clifford, 2004). Though the author reported high correlations between ECERS-R assessments of the same classrooms over time in the NCEDL studies, all reported numbers came from personal communications between the author and researchers working on the NCEDL projects, as opposed to reports of correlations in the original study results. If the ECERS lacks temporal stability, depending on the time that a classroom was assessed, it may or may not reach licensing standards or reflect its highest quality so that further reimbursements can be earned. Further research is needed to determine whether the ECERS does indeed yield a reliable measure of global environmental quality across time in the same classroom.

Although an instrument does not need to be valid to be reliable, reliability is a necessary (but not sufficient) requirement for validity (Crano & Brewer, 2002). The validity of a scale generally refers to "the extent of correspondence between variations in the scores on the instrument and variation among respondents on the underlying construct being studied" (Crano & Brewer, p. 45). Similar to reliability, there are several types of validity that are important to examine in regards to the value of a scale's use. This section focuses specifically on the predictive validity of the ECERS. Instruments like the ECERS involve considerable money spent on conducting the assessments and reimbursing centers at higher rates. These activities beg the question, however, of whether or not the items being measured by the ECERS have any impact on child outcomes.

The predictive value of the ECERS has already been discussed in the section involving longitudinal child variables. Although many of the studies described in this paper used the ECERS to measure quality, this section will not re-review studies that either used the ECERS in combination with other measures or used the ECERS without including all of the instrument's items. Studies utilizing the ECERS in a way other than that which was intended by the developers do not allow the individual contribution of the entire ECERS measurement to prediction of child outcomes to be assessed.

Only one study reported in this paper that used quality to predict children's longitudinal academic and/or social outcomes used the ECERS-R as it was intended, including all subscales and used in classrooms serving children no younger than two and a half years of age, and did not combine it with other quality measures (Sylva et al. 2006). Researchers in this study did report that, after controlling for child age, pretest scores, and child and family background characteristics, the total ECERS-R from age three was not predictive of children's academic achievement at age five, but did significantly predict some of the teacher ratings of children's social behavior at age five. Due to the lack of studies examining the predictive nature of preschool quality for children's academic and social outcomes that do use the ECERS in its entirety and independently of other measures of quality, it is difficult to ascertain the predictive validity of the instrument. Further research is needed to determine the predictive value of the ECERS, especially considering the scope of the instrument's current use in policy matters. If the ECERS is predictive of children's outcomes, the temporal reliability of the measure becomes particularly important. Different scores within the same year for a classroom might lead one to assume that, if the ECERS-R is truly predictive of child outcomes, quality assessed at one time period might reveal that predictive nature while quality assessed at another time might not.

Conclusions and Hypotheses

The current focus on the quality of early childhood care spans the research community, the educational community, and the general public. People are concerned with the experiences to which the youngest members of the community are being exposed. The number of children enrolled in preschool programs is continuing to rise, and it is more important than ever to ensure that those programs are of high quality.

However, quality assurance becomes difficult when the definition of quality is unclear. The variety of quality definitions, stemming from a lack of consensus as to what constitutes quality, makes it hard to assess whether early care environments are offering high-quality experiences to young children. In addition, there exists a wide variety of instruments currently being used to measure child care quality. The picture of quality in a classroom is entirely dependent on the focus of the instrument used for quality assessment.

Because of the different instruments utilized, research has found contradictory evidence that speaks to the relationship between quality and child- and teacher-level variables. Although it is assumed that higher quality leads to better child outcomes both academically and socially, research on this topic has shown mixed results. These mixed findings may be the result of not measuring the right aspects of an environment, or not measuring those aspects in the right way.

Additionally, research on classroom quality carries with it several design difficulties that have not often been considered in existing studies. Much of the research on quality focuses on immediately or contemporaneously measured child variables. This research often implies that quality causes differences in child achievement. However, the majority of such studies do not look at the prediction of quality from one year to outcomes in another year. In addition, researchers often ignore the initial achievement of the children. This type of research can only lead to implications about the co-occurrence of quality and achievement as opposed to a more direct causal relationship. Another difficulty often ignored is the appropriate analysis to use when many target children are nested within classrooms. Simple correlation and regression analyses are not appropriate when the independence of units cannot be assumed. A final and serious difficulty arises when the results of studies of quality are compared to one another. Because different instruments and different methods of using the same instruments are used to assess quality, comparisons across studies are not often valid.

The main assessment measure currently being used to look at early child care quality is the ECERS. The use of the ECERS is widespread, despite contradictory findings from research as to the link between ECERS scores and child outcomes, as well as serious issues concerning the psychometric properties of the instrument. Despite the plethora of research using the ECERS is a quality assessment tool, it is still unclear as to whether the ECERS is a temporally reliable or predictively valid measure. The ECERS is often tied to state licensing standards, effectively linking scores on the ECERS to center reimbursements. When an instrument such as the ECERS is affecting the livelihood of public citizens, it is essential that the instrument has been shown to be a reliable and valid measure. Not only have the limited psychometric analyses of the ECERS not revealed promising results to warrant its widespread and influential use, but the method of scoring the instrument also raises serious concerns about how it is currently being used. Given that the ECERS is currently used to assess quality perhaps more than any other measure, both domestically and internationally, and is often linked to child care program funding, the research community is obligated to determine whether the ECERS is offering an accurate picture of quality.

This study sought to address four main research questions. The first group involved examining the consequences of using alternative scoring methods on a classroom's resulting quality score. The second group of questions looked at the extent of agreement among experts in the field about whether the items included in the ECERS-R reflect components thought to be essential to quality. The third group of questions concerned the psychometrics of the instrument with alternative scoring methods and item inclusion. The final group of questions attempted to make conclusions about which scoring method and make-up would best alleviate some of the policy-level concerns that the use of the ECERS-R poses. The general research questions were:

1. To what extent are the results of the ECERS-R affected by the scoring conventions that are currently used by the developers and the two alternative scoring methods used by this researcher?
2. To what extent do field experts agree among themselves and with the instrument's developers on the organization and content of the ECERS-R?
3. What are the psychometric properties of the original ECERS-R and those of a new version of the ECERS-R based on expert opinion?

4. What are the policy implications associated with using a new version of the ECERS-R that is created based on aspects of a classroom that field experts consider important to quality?

CHAPTER III

RESEARCH DESIGN AND PROCEDURES

Introduction

The purpose of this study was to examine an instrument used to measure quality in early childhood settings that is currently involved in high-stakes policy-related issues and to examine what the properties of this instrument might reveal about conclusions drawn from the scores it produces. Specifically, this study used the ECERS-R, the most widely-used measure of early childcare quality, to investigate how current uses of quality assessments might be leading to conclusions about the quality of care a child experiences and, as a result, to issues concerning the links between center quality scores and state policy.

Research Design

This study consisted of secondary analyses of data available from classroom observations in three separate studies of prekindergarten classrooms, as well as original data collected from field experts for the purpose of this study.

Research Instruments

Early Childhood Environment Rating Scale-Revised Edition (ECERS-R)

The ECERS-R (Harms, Clifford, & Cryer, 1998), an instrument used to rate the global quality of early childhood educational environments, is comprised of 470 individual indicators, organized into 43 items under seven subscales: Space and Furnishings, Personal Care, Language-Reasoning, Activities, Interaction, Program Structure, and Parents and Staff. As discussed earlier in this paper, the ECERS-R is the revision of the original ECERS instrument. The ECERS-R is a commercially available instrument with available video training tapes and an extensive guide that describes the items in detail (Cryer, Harms, & Riley, 2003). Many studies only use and/or report data from the first six subscales, omitting the subscale pertaining solely to adults. The instrument is used primarily to assess the quality in classrooms serving children from 2.5 to 5 years of age.

Each of the 43 items has individual indicators that are used to determine a classroom's score on that item. The indicators are grouped under four scale points, 1, 3, 5, and 7, representing anchors of Inadequate, Minimal, Good, and Excellent. Each indicator, beginning with the ones under the Inadequate anchor, are scored as present or not by the observer. In the traditional scoring method of the instrument, if all of the indicators under an anchor are marked as present (or positive), the indicators under the next anchor are then scored. This method continues until an indicator is scored as not present (or negative). Once a negative indicator is reached, the rest of the indicators under that anchor are scored, and then the observer stops scoring that item at that anchor; higher anchored ratings on the item are not scored. Further detail about the stop-rule is given in the section entitled "ECERS-R Stop-Rule" below.

All of the indicators under ECERS-R anchors are positively worded except for the indicators under the “Inadequate” anchor, which are always negatively worded. Checking “yes” for any of the indicators under “1” is equivalent to checking “no” for any of the indicators under “3”, “5”, or “7”, and leads to discontinuing scoring the item. A classroom receives a score on an item based on the anchor under which the first negatively-scored indicator appears. For example, if all of the indicators under “1” are absent (which is positive) and all of the indicators under “3” are present (which is positive) but only half of the indicators under “5” are present, a classroom cannot receive a higher score than a 4 for that item. Item scores under each subscale are averaged for subscale scores and, ultimately, total quality scores that range on an interval scale from one to seven. Table 2 displays the organization of the ECERS-R.

Table 2

ECERS-R Layout

Sample ECERS-R Item						
1	2	3	4	5	6	7
(Anchor: Inadequate)		(Anchor: Minimal)		(Anchor: Good)		(Anchor: Excellent)
Y N *Indicator*		Y N *Indicator*		Y N *Indicator*		Y N *Indicator*
Y N *Indicator*		Y N *Indicator*		Y N *Indicator*		Y N *Indicator*
Y N *Indicator*		Y N *Indicator*		Y N *Indicator*		Y N *Indicator*
Y N *Indicator*		Y N *Indicator*		Y N *Indicator*		Y N *Indicator*

The developers of the ECERS-R recommend an observation period of at least three hours, followed by a brief teacher interview to answer questions about items that were not observed. According to the ECERS-R manual, the developers used two observers in 21 classrooms to calculate inter-rater reliability for the instrument. At the indicator level, the authors report 86.1% agreement. At the item level, the authors report 48% exact agreement between raters and 71% agreement within-one. The authors also report correlations between the two observer scores as .921 (Pearson) and .865 (Spearman). Although the method of calculating inter-rater reliability is not described in detail, an observer agreement form is available for download from the developers' website (*Environment Rating Scales*, n.d). This form allows two observers in the same classroom to first come to a consensus score and then calculate their agreement with the consensus score, instead of the more customary system of comparing their raw scores to each other and calculating agreements and disagreements. In addition, the form does not provide space to calculate exact agreement but only leaves space for within-one point calculations. Internal consistency was also reported by the developers in the instrument manual at both the subscale and total score levels. Regarding consistency at the subscale level, consistency scores ranged from .71 to .88, and the internal consistency score for the entire measure was .92.

ECERS-R Stop-Rule. Once an observer negatively scores an ECERS-R indicator, the observer must finish all of the indicators under that anchored point and then stop scoring that item. Anchored descriptions are for odd-numbered ratings only (1, 3, 5, 7). If at least half of the indicators under an anchor are marked positively, the observer gives the item the even numbered rating just before the anchored point. If fewer than half of the anchor's indicators are marked positively, the classroom must receive a score of on the next lowest odd numbered rating (i.e., dropping back two points). The consequence of this stop rule is that if a classroom receives a score of "3" on an item, it can be assumed that each indicator below that anchor was positively scored. However, several of the indicators above the "3" rating may have actually been present but not scored due to the stop-rule. The stop-rule used by ECERS-R raters is of primary importance in the proposed study. In most studies using the ECERS-R, because of the stop-rule, any indicator under an anchored point above where the stop rule came into effect would not be scored during that observation period, even though the behavior might have been scored positively had the observer continued scoring that item.

Indicator Survey

The *Indicator Survey* was constructed to allow respondents to evaluate the organization and representation of quality dimensions in the ECERS-R. The survey was an electronic spreadsheet document that listed each indicator and asked respondents to indicate in which of the ECERS-R dimensions the indicator best fit. These responses could then be checked against where the ECERS-R developers had placed the indicator, providing a validation of each individual indicator as well as the organization of the scale as a whole. In addition, the survey contributed to an investigation of the construct validity of the instrument. Experts were asked how much each indicator was important to their own personal definitions of quality. Thus, a respondent might agree that an item belonged in a certain dimension, agreeing with the ECERS-R placement, but not agree that the item was important to the respondent's definition of quality.

In the process of survey development, all of the indicators from the first six subscales (all of the subscales directly involving the child's experience, excluding the seventh subscale that involves the parent and staff provisions only) were first listed in the order that they appear in the ECERS-R, which totaled 397 indicators. Because the indicators were not designed to stand alone without their corresponding item and subscale names, some indicators when read in isolation were unclear as to the specific piece of the environment to which it is referring. Each of the indicator's wording was supplemented based on the name of the item if the indicator was not clear. For instance, under the Space and Furnishings subscale, the Indoor Space item has an indicator which originally reads "Insufficient space for children, adults, and furnishings." This indicator was included in the survey as "Insufficient *indoor* space for children, adults, and furnishings" to make it clearer to the reader. The order of the list of indicators was then randomized before being put into the survey so that indicators were not already grouped according to the ECERS-R organization when experts were asked to place them into categories.

For each indicator listed on the survey, the respondent was asked to indicate the extent to which he or she agreed that that indicator was an important quality component of each of the first six ECERS-R subscales. The respondent could check the alternatives of "A great deal", "Somewhat", or "Not at all" for each of the 6 subscales. In addition, the respondent was asked to indicate the extent to which he or she felt that each indicator was important to his/her personal definition of quality, using the same three alternatives as the previous question.

At the conclusion of the rating portion of the survey, there were several open-ended questions the respondent was asked to answer. The questions allowed respondents an opportunity to express whether there were important indicators of quality that were not listed on the survey, whether they felt any of the given indicators represented constructs other than classroom quality, and how they saw the relative importance and definitions of each of the ECERS-R subscales. The survey instructions and a sample page are provided in Appendix A. Scoring of the instrument is described in the Analyses section.

Research Sites and Participants

Secondary Data Sets

Three groups of previously-collected data were used in this study. The first set of data was comprised of ECERS-R scores from 118 Pre-Kindergarten classrooms collected as part of an Early Math Project. Of those 118 classrooms, 70 classes were located in Tennessee and 48 were located in California. The second set was comprised of ECERS-R scores from 122 classrooms in Missouri collected as part of a Quality Rating System (QRS) pilot study. The third data set was comprised of ECERS-R scores from 21 classrooms in Tennessee collected as part of the Preschool Curriculum Evaluation Research (PCER) grant project. Each of these data sets is described in more detail in the following sections.

Group 1: Early Math Project Classrooms. The Tennessee and California classroom data were collected as part of the Early Math Project and included classrooms in both the state-funded public school system and Head Start program. All classrooms in Tennessee and California served primarily low-income children.

The Early Math Project was a four-year randomized control trial funded by the Institute of Education Sciences evaluating the effectiveness of an early math curriculum intervention on enhancing children’s math knowledge and school achievement. The principal investigators for the primary award for this project were Prentice Starkey and Alice Klein of the University of California at Berkeley with a subcontract to Dale Farran and Mark Lipsey at Vanderbilt University (IES award number R305K05186). Vanderbilt handled the training and scoring of the ECERS-R at both the California and Tennessee sites. Both California and Tennessee sites enrolled classrooms connected to Head Start and to the public schools. There were 32 classrooms in Tennessee and 20 classrooms in California housed in public schools; these classrooms served prekindergarten four year olds who were from low income families but who were not diagnosed with a special need. Within the Tennessee site 38 classrooms were housed in Head Start centers; 28 California classrooms were Head Start. In the original study, sites were randomly assigned to conditions. Data for this dissertation were collected as part of the larger study to gather information on classroom quality before the intervention training began.

The total number of classrooms was 70 in Tennessee (32 public school classrooms and 38 Head Start classrooms) and 48 in California (20 public school classrooms and 28 Head Start classrooms).   The majority of Tennessee and California classrooms were only observed one time and were observed during the fall/winter. Nine of the Tennessee classrooms were observed in the spring instead of the fall because there was not a permanent lead teacher in the classrooms until the spring. Four of the Head Start classrooms in Tennessee were observed twice because there was a change in the lead teacher. The second observations in these four classrooms were done in the spring. Because the nature of the analyses in this dissertation did not require independence of classroom observations, all observations were included, for a total of 122 observations. The ECERS-R was used in the original study to gather information about the quality of classrooms and to compare their ratings to others in published studies. In all of the classrooms in this group, the ECERS-R was scored without observing the stop-rules and every indicator was scored.

Group 2: Quality Rating Systems Project Classrooms. In 2003, the Midwest Child Care Research Consortium (MCCRC), an organization combining representatives from universities and state-funded agencies, initiated a study examining the development and implementation of rating systems for classroom quality. In Missouri, one of the states in the Consortium, the goals were to provide the public with more information about the quality of child care facilities in their areas, increase accountability at the level of policy and funding sources, and give programs feedback so that they could improve the quality of their own classrooms. The Missouri classroom data were collected as part of the Quality Rating Systems pilot study at the University of Missouri-Columbia. The principal investigator for the Missouri portion of this study was Kathy Thornburg. Forty-one counties in the state were selected to participate (33 rural and 8 urban). From the rural counties, because of the limited number of programs in rural areas, all licensed child care programs in those areas were invited to participate in the study. In the urban counties, licensed child care programs were randomly selected and invited. A final sample of 122 classrooms in 70 programs was scored using the ECERS-R. Ninety percent of the programs received child care subsidies, and 37.1% were accredited by the National Association for the Education of Young Children (National Association for the Education of Young Children, n.d.). Of the 122 classrooms, 83 were observed only one time and 39 were observed twice in one year. Data from all observations were used in this study's analyses. All 122 classrooms were observed using the ECERS-R without observing the stop-rules, meaning every indicator was scored. A total of 161 observations were included in analyses.

Group 3: Preschool Curriculum Evaluation Research Project Classrooms. The third set of data was comprised of ECERS-R scores from 21 Pre-Kindergarten classrooms in public schools in Tennessee. These classrooms were involved in the Preschool Curriculum Evaluation Research (PCER) grant project. The Tennessee PCER project began in 2002, was one of seven projects funded nationally under the PCER initiative, and was headed by principal investigators Dale Farran and Mark Lipsey. The project was a four-year randomized control trial designed to evaluate the effectiveness of two different preschool curricula. All Tennessee classrooms were located in rural counties in the state and primarily served students from low-income households.

The ECERS-R observations were mandated by the national evaluation; classrooms observations were conducted by employees of the Research Triangle Institute (RTI). Each of the 21 classrooms was observed twice within the same year, once in the fall and once in the spring. All classrooms were scored with the ECERS-R using the traditional stop-rule scoring method. Classrooms in this data set were not the same classrooms in the Tennessee PCER data set. The details about each of the data sets are displayed in Table 3.

Table 3

Secondary Data Samples

Number of Classrooms	State	Original Project	Number of Years Observed	Number of Times Observed Each Year	Scoring Rules	Observations Contributed
70	TN	Early Math	1	1	No stop-rule	70
4	TN	Early Math	1	2	No stop-rule	4
48	CA	Early Math	1	1	No stop-rule	48
83	MO	QRS	1	1	No stop-rule	83
39	MO	QRS	1	2	No stop-rule	78
21	TN	PCER	1	2	Stop-rule	42

Expert Panel

For the additional data set, 16 experts in the field of child development and/or early childhood education were selected. Experts were nominated by the researcher's dissertation committee and represented researchers in the field who are affiliated with an academic institution in the United States but who are also still involved with early childhood classrooms so that they have a deep understanding of both research and practice.

Each nominated expert was sent a letter requesting their participation in this study. Thirty-four experts were originally sent request letters. Of those 34 experts, 22 agreed to participate. Of those agreeing to participate in the study, 16 returned their completed survey forms. Of the respondents who completed the survey, 14 had PhDs and 2 had EdDs. The majority of respondents had obtained their doctorate degrees in psychology, child development, early childhood education, or a related field. There were also respondents with degrees in early childhood special education and educational administration. The experts indicated their areas of specialization to include early literacy and language, self-regulation, play, teacher talk, quality measurement, school readiness, early educational policy, early science education, alignment of assessment to standards, social development, emotional development, and professional development. Eight of the 16 respondents had taught in an early childhood classroom, 9 had experience as a director and/or supervisor, and all 16 had taken on the role of researcher in early childhood education. Twelve of the respondents had used the ECERS-R before in some capacity, and only 2 had not used either the ECERS-R or any other measure to assess the quality of a classroom.

Procedures

Secondary Data Sets

Early Math Project data: California and Tennessee. As in many other studies using the ECERS-R, only 37 of the 43 items were scored, omitting the questions concerning parents and staff, as the first 37 items are considered to be child and program-related. All observers had over 60% inter-rater agreement on exact item scores, over 78% agreement on item scores within one, and over 88% agreement on indicator scores. Observers went into each classroom and observed for the entire length of the school day, a period longer than the minimum recommended by the instrument's developers (Harms, Clifford, & Cryer, 1998), in order to see actual instances of as many of the items on the scale as possible. (Items on the ECERS-R can be scored by teacher report in an interview following the observation if the observer did not have the opportunity to see instances of an event, e.g. nap procedures. To have as few items scores by teacher report as possible, the observers in this study stayed for a longer period of time.) A total of 118 classrooms were observed, 48 in public schools and 70 in Head Start centers. Because 4 classrooms were observed twice with different lead teachers, a total of 122 observations were included in analyses.

Quality Rating Systems pilot study: Missouri. All 43 items of the instrument were scored, but to be consistent with the other data sets, only the first six subscales were used in these analyses. All observers met the study's requirement of 85% inter-rater agreement on exact item scores and 90% agreement on item scores within one. Additionally, reliability checks were done on each observer with the same requirements for every tenth observation conducted, or after six months of their last check, whichever came first. In this study, all observers met reliability requirements at each check point. Observers were in each classroom for a period of three to four hours.

PCER grant data: Tennessee. Collection of ECERS-R data was conducted by employees of the Research Triangle Institute. According to the national report on the PCER project, in the preschool year, RTI data collection staff were trained to conduct the classroom observations and teacher interviews. Only the first six subscales of the ECERS-R were scored. RTI recruited classroom observers who had a background in early childhood education and previous experience using the ECERS-R measure. Thirteen observers were recruited for the fall data collection. All trained staff participated in two additional practice days during which training reliability was achieved in the week following a group training session. Classroom observers with limited observation experience participated in two additional days of practice in classroom settings. No more detailed information on reliability is provided by RTI or the Institute for Education Sciences, the funding agency for the national study.

Expert Panel

After permission was obtained either through e-mail or by phone, respondents were e-mailed a Microsoft Excel file listing individual indicators from the Early Childhood Environment Rating Scale. Respondents were asked to rank the indicators within the given categories, rank the indicators as to their representation of quality, and answer certain open-ended questions about their individual sorting and ranking decisions. Additionally, respondents were requested to complete the survey within a six-week time period. Panel members submitted their completed survey forms via e-mail or mailed hard copy, at which time the forms were assigned an anonymous identification number. A sample page of the survey can be found in Appendix A. After the researcher received the completed survey, each respondent was offered a check for $100 in appreciation of their contributed time and effort.

Missing Data

There were several classroom observations from the secondary data sets that had missing data. Of the 325 classroom observations that were scored at least at the item-level, there were no missing data at that level. Of the 283 observations that were scored at the Indicator-level, 280 were scored at the Indicator-level up to the stop-rule. One observation had missing indicator-level data for an entire item, and a total of 8 indicators were missing from 6 additional items. For the case where the entire item was missing, the case was given a score based on the maximum number of indicators possible given their item-level score (which was not missing). For the missing individual indicators, surrounding indicators were examined together with the corresponding item-level score and the indicator was scored accordingly. For example, if a classroom had a score of “4” on an item but had an indicator with missing data under the anchor of “5”, the other indicators under the “5” anchor would be examined. If less than half of those indicators had been positively scored, the indicator with missing data must have been positively scored also for an item score of “4” to be given. Of 281 observations with enough data to calculate sums of indicators, 13 observations had information missing for more than 40 indicators, resulting in 268 observations that were scored at the Indicator-level with non-missing data for at least 90% of the indicators. For cases with information missing for less than 10% of the indicators, data were imputed from the modal rating of the rest of the sample for each indicator. The 13 observations that did not have enough data to be scored at the indicator-level were included in analyses that utilized the alternative scoring methods described more in depth in the next section.

Regarding the Indicator Survey, some of the indicators on all but three of the expert surveys had missing data. Across all 16 surveys, 7 experts had missing subscale placement data for 1 to 4 indicators with a mean number of indicators with missing placement data for all experts of .94 (SD=1.23). Across all 16 surveys, 13 experts had missing data on the importance of indicators to their personal definitions of quality for 1 to 6 indicators with a mean number of indicators with missing importance data for all experts of 2.06 (SD=1.77).

Research Questions

Research Question 1: To what extent are the results of the ECERS-R affected by the scoring conventions that are currently used by the developers and the two alternative scoring methods used by this researcher?

In order to answer the first research question concerning the consequences of using alternative scoring methods on ECERS-R data, scores from three different scoring methods were used. 1a) Method 1 (traditional method) involved the ECERS-R scoring procedure in which indicators were only scored until the stop-rule was necessitated and the total score was an average of the item scores. 1b) Method 2 (summative-no stop method) involved a summative scoring procedure in which every indicator was scored (the stop-rule was not used) and the total score was a sum total of all of the positively-scored indicators. 1c) Method 3 (summative-stop method) involved a summative scoring procedure in which indicators were only scored until the stop-rule was necessitated and the total score was a sum total of all of the positively-scored indicators. In both of the summative methods, the negatively-worded indicators on the ECERS-R under the “Inadequate” anchor were recoded so that a score of “0” on one of those indicators would mean that a classroom was counted negatively on that indicator.

Because both of the alternative summative methods involve a total score, the totals are affected by those ECERS-R indicators that may be scored as “Not Applicable” (NA) instead of receiving a 1 or 0. For instance Item 1, Indoor Space, has an indicator under the Minimal anchor regarding whether the space is accessible to children or adults who may have physical disabilities. If a classroom does not have any children or adults who meet this description, that indicator may be scored NA. If that classroom got all other indicators scored positively, they would receive a total item score of 13. Another classroom that did include a child with a physical handicap but also got all indicators scored positively would receive a total item score of 14. However, using the traditional method, both classrooms would receive an item score of 7. There are 50 indicators in the ECERS-R that have NA as an allowed scoring option. Because it was unclear whether these indicators would affect the outcome of the analyses in this study, both of the summative methods were included in initial analyses with and without NA items removed from the scores.

Method 1 compared to Method 2. When comparing the traditional method and the summative-no stop method, data from every time point of those classrooms in which every indicator was scored was used. Because independence of classrooms did not need to be assumed, classrooms that contributed more than one observation were included for a total of 268 observation points. First, each classroom was given a score using the traditional scoring method of the instrument, which bases a classroom's total score on the average of item scores. Based on those scores, each classroom was ranked in order from lowest to highest and received a number according to its rank (1 being the lowest and 268 being the highest). Second, a different scoring method was applied to each classroom. Each classroom was given a score based on the total number of indicators that were positively scored when every indicator was included in the scoring process. Based on those scores, each classroom was ranked in order from lowest to highest and received a number according to its rank (1 being the lowest and 268 being the highest).

A correlational analysis was run on the rank orders of the classrooms based on the average scoring method and the summative-no stop method. The resulting correlation spoke to the difference between using a traditional scoring method and using a new scoring method that differs in both scoring algorithm and stop-rule use. A classroom might have had a higher or lower standing in relation to other classrooms when the method was altered than they would if the traditional method had been used.

Method 1 compared to Method 3. When comparing the traditional method and the summative-stop method, the alternative method that involved observations where indicators were scored up to the stop-rule and the total score was determined by the sum of all positive-scored indicators, data from all observed classrooms at all time points were used. Again, this analysis did not need to assume independence of classrooms because each classroom's score was compared to itself using alternative scoring methods. Therefore, data from all observed classrooms were used, including multiple observations from the same classrooms. For those classrooms in which every indicator was scored, it was possible to go back and rescore that classroom with the stop-rule in place using the summative-stop method by removing those indicators from the total summed score that would not have been scored if the stop-rule had been in place at the time of the observation. A total of 281 observation points were used.

First, each classroom was given a score using the traditional scoring method of the instrument, which bases a classroom's total score on the average of item scores. Based on those scores, each classroom was ranked in order from lowest to highest and given a number according to its rank (1 being the lowest and 281 being the highest). Second, a different scoring method was applied to each of the 281 classroom. While still observing the stop-rule, rather than using the average score (on a scale of 1 to 7) based on all 37 items, a score was given based on the total number of indicators scored positively. An example of how this was done is given in Table 4.

As depicted in Table 4, in this scenario the observer scored all of the indicators under "1" and "3" positively (indicated by "No" under "1" and "Yes" under "3" because of the negatively-worded indicators under the “1” Anchor) but only two of the four indicators under “5” positively. Using the traditional scoring method, this classroom would receive a score of “4” on this item. Using the summative-stop method, this classroom would receive a “10” on this item (the sum total of all indicators scored positively). Each classroom was given a total ECERS-R score based on this summative-stop method. Based on those scores, each classroom was ranked in order from lowest to highest and received a number according to its rank (1 being the lowest and 281 being the highest). A correlational analysis was run on the rank orders of the classrooms based on the traditional scoring method and the summative scoring method. The resulting correlation spoke to the difference between using the two scoring methods, traditional and summative-stop, with items scored. A classroom might have had a higher or lower standing in relation to other classrooms when using the traditional method than it would have if the summative-stop method had been used.

Table 4

Example of Traditional v. Summative Scoring Using Stop-Rule

Sample ECERS-R Item						
1	2	3	4	5	6	7
(Anchor: Inadequate)		(Anchor: Minimal)		(Anchor: Good)		(Anchor: Excellent)
N *Indicator*		Y *Indicator*		Y *Indicator*		Y N *Indicator*
N *Indicator*		Y *Indicator*		Y *Indicator*		Y N *Indicator*
N *Indicator*		Y *Indicator*		N *Indicator*		Y N *Indicator*
N *Indicator*		Y *Indicator*		N *Indicator*		Y N *Indicator*

Method 2 compared to Method 3. When comparing the summative-no stop and the summative-stop methods, data from every time point of those classrooms in which every indicator was scored were used. Again, because independence of classrooms did not need to be assumed, all classrooms were used including those that contributed more than on observation for a total of 268 observation points. The procedure involved, first, each classroom was given a score based on the total number of indicators that were positively scored while observing the stop-rules. Based on those scores, each classroom was ranked in order from lowest to highest and received a number according to its rank (1 being the lowest and 268 being the highest). Second, each classroom was given a score based on the total number of indicators that were positively scored when every indicator was included in the scoring process. Based on those scores, each classroom was ranked in order from lowest to highest and received a number according to its rank (1 being the lowest and 268 being the highest). An example of how this was done is given in Table 5.

As depicted in Table 5, if the observer had been using the stop-rules, he/she would not have scored any of the indicators under the "Excellent" anchor. Therefore, the classroom would have a summative score of "9" (the number of indicators scored positively that would have been observed using the stop-rule scoring method). If the stop-rule had not been in place, the observer would have scored every indicator regardless of how many were negatively scored, and the classroom would have a summative score of "11." A correlational analysis was run on the rank orders of the classrooms based on the summative-stop method and the summative-no stop method. The resulting correlation spoke to the difference between using and not using the stop-rule. A classroom might have had a higher or lower standing in relation to other classrooms when using the stop-rule than they would have had the stop-rule had not been not used.

Table 5

Example of Summative-No Stop v. Summative-Stop Scoring

Sample ECERS-R Item						
1	2	3	4	5	6	7
(Anchor: Inadequate)		(Anchor: Minimal)		(Anchor: Good)		(Anchor: Excellent)
N *Indicator*		Y *Indicator*		Y *Indicator*		N *Indicator*
N *Indicator*		Y *Indicator*		N *Indicator*		Y *Indicator*
N *Indicator*		Y *Indicator*		N *Indicator*		N *Indicator*
N *Indicator*		Y *Indicator*		N *Indicator*		Y *Indicator*

Policy implications of scoring method. In addition to the comparison of relative ranking, the researcher also looked at changes in funding status if scoring methods were altered. Each state that uses the ECERS-R as part of its evaluation of early childhood programs typically has a minimum score required for reimbursement rates to be distributed to each program. Because every state has different requirements, Tennessee, a state in which much of this study's data was collected, will be used as the example for analytical purposes in this paper. In Tennessee, in order to quality for the state's Star Quality program, as mentioned in Chapter 2 of this paper, a center must have obtained an overall average ECERS-R score of a 4.0 or higher. A classroom's score, whether measured with Methods 1, 2, or 3, was determined to be either above or below that cut-off.

Regarding the traditional scoring method, a classroom's total average score ranged from 1 to 7. Comparing a classroom's total average score to the cut-off score of "4.00", each classroom was given a 1 (at or above the cut-off) or a 0 (below the cut-off). Regarding the summative-stop and summative no-stop methods, the number of indicators under each item that would be required to be positively scored in order to receive a score of "4.00" were summed, yielding a total number of indicators for the instrument that would be necessary to receive a “4” overall. For example, in Table 5 above, an item score of 10 would be necessary to receive a score of “4” (counting 4 indicators under “Inadequate”, 4 indicators under “Minimal”, and 2 of the 4 indicators under “Good”). The number of indicators required to receive a score of “4” on each item were summed across all 37 items. For the summative methods with NA indicators included, the cut-off score was 253; without NAs it was 227. Comparing a classroom’s total summative-stop and summative-no stop scores to the total number of indicators necessary to receive an overall “4.00”, each classroom was given a 1 (above the cut-off) or a 0 (below the cut-off).

Because the nature of the data consisted of three comparisons of paired data in which the dependent variable is dichotomous, a McNemar two-tailed nonparametric test was used to examine whether the extent of change in 1’s from using one method to another, in either direction, was significant. This test was run three different times to compare each method to each of the others. As in the first analyses regarding relative ranking, 268 classrooms were used to compare Method 1 to Method 2 and Method 2 to Method 3, and 281 classrooms were used to compare Method 1 to Method 3.

Classroom characterization by alternate scoring methods. A third comparison of scoring methods, in addition to classroom ranking relative to each other and relative to a funding cut-score, is the comparison of a classroom’s rank relative to the ECERS-R quality categories. As mentioned earlier in this chapter, the traditional scoring scale on each of the ECERS-R items and the total measure score ranges from 1 to 7. Within that scale, the indicators are grouped under four scale points, 1, 3, 5, and 7, representing anchors of Inadequate, Minimal, Good, and Excellent. Therefore, the total numerical score that a classroom receives when scored traditionally can be equated with one of these anchors. In this study, scores in-between anchors were equated with the closest anchor. For example, a score of “5.00” was translated as “Good”, while a score of “1.68” was translated as “Inadequate”, and a score of “6.15” was translated as “Excellent.” In cases where there was no closest anchor, scores were equated with the lower anchor. For example, a score of “6.00” was translated as “Good” rather than “Excellent, whereas a score of “6.01” was translated as “Excellent.”

Using the traditional scoring method, a classroom’s total average score was equated with one of the four anchor categories. Each classroom was given a categorical score of 1 to 4 depending on the category represented by their total score. Scores between 1.0000 and 2.0000 were given a 1, scores between 2.001 and 4.000 were given a 2, scores between 4.0001 and 6.0000 were given a 3, and scores above 6.0000 were given a 4.

Using the summative-stop and summative-no stop scoring methods, the number of indicators under each category anchor of each item were summed, yielding the total number of possible indicators that corresponded to an overall score equated with each of the four categories. Comparing a classroom’s total summative score to the total number of indicators necessary to receive an overall score in each category, each classroom was given a categorical score of 1 to 4 depending on the category represented by their total score. For the summative methods with NAs included, scores between 0 and 145 were given a 1, scores between 146 and 252 were given a 2, scores between 253 and 344 were given a 3, and scores between 345 and 397 were given a 4. For the summative methods without NAs included, scores between 0 and 130 were given a 1, scores between 131 and 226 were given a 2, scores between 227 and 311 were given a 3, and scores between 312 and 397 were given a 4.

Because the nature of the data consisted of three comparisons of paired data in which the dependent variable was categorical, a Wilcoxon matched pairs signed-ranks two-tailed nonparametric test was used to examine whether the extent of change in category from using one method to another, in either direction, was significant. This test was run three different times to compare each method to each of the others. As in the previous comparisons, 268 classrooms were used to compare Method 1 to Method 2 and Method 2 to Method 3, and 281 classrooms were used to compare Method 1 to Method 3. In addition, the type of category changes resulting from a method difference was examined. The total number of classrooms that began in each category and changed to each of the other categories when the method was changed was examined.

Research Question 2: To what extent do field experts agree among themselves and with the instrument's developers on the organization and content of the ECERS-R?

Regardless of the scoring method, a more general question about the ECERS-R is one of construct validity, that is whether the most important classroom characteristics are being examined and whether those characteristics are correctly operationalized under the domains of the instrument. Two types of analyses were conducted on the *Indicator Survey* to address this question, one involving the subscale organization of the instrument and another involving the extent of agreement about ECERS-R content.

Respondents were asked to first indicate their agreement about whether each indicator belonged under each subscale. Respondents could indicate a great deal of agreement, some agreement, or disagreement for each indicator and each subscale. Respondents were also asked to indicate their agreement about the importance of each indicator to their personal definitions of quality, using the same three response options. The question involving subscale organization looked at the extent to which field experts agreed with ECERS-R developers on the placement of indicators under certain subscales. In addition, the extent of agreement among field experts on the organization was examined. The question involving the content of the ECERS-R looked at the extent to which field experts agreed that ECERS-R indicators were important components of childcare quality.

Agreement on organization. For the subscale analysis, respondents scored each indicator with a “1” or a “0” under each domain (or subscale). A score of “1” corresponded to the domain under which the respondent agreed the indicator should be placed, regardless of the degree of agreement. The agreement levels of “A great deal” and “Somewhat” were collapsed into one category representing agreement. If the respondent checked “A great deal” or “Somewhat” for more than one domain for a given indicator, each of those domains received a score of “1.” This process resulted in a record for each domain that listed the indicators placed there by the respondent. Because the indicators were taken directly from the ECERS-R, the researcher was able to create a list for each domain that gave the indicators placed there by the instrument’s developers. Those two lists were compared to each other. Each respondent received a percent agreement indicating the portion of indicators that the respondent placed under the same domain as the ECERS-R developers. Percentage agreement among respondents was also calculated by looking at the number of indicators that all experts placed under the same domain.

Additionally, by summing the “1’s” under each domain for each indicator across respondents, every indicator received a score corresponding to the number of respondents that agreed that the indicator applied to each particular domain. These sums were compared to the domain that the ECERS-R developers placed the indicator under. The sum for the domain that the indicator was taken from in the ECERS-R was divided by the total number of respondents, yielding the percentage of respondents who agreed with the ECERS-R developers on placement of each indicator. Means and standard deviations were calculated across indicators, revealing the number of indicators on which the respondents and the ECERS-R developers agreed on concerning subscale placement.

Concerning the analysis of the extent of agreement, the same statistics were calculated (percent agreement between each expert and the ECERS-R developers, percent agreement among experts, and percent of experts that agree with the ECERS-R developers) at the highest level of agreement only ("A great deal"). This set of analyses assigned a "1" for each indicator under each subscale that the expert placed the indicator with a great deal of agreement.

Agreement on content. In addition to agreement between experts and ECERS-R developers concerning organization, the extent to which experts agreed with developers on the content of the ECERS-R was also examined. The study looked at how well the operationalization of quality from the ECERS-R matched the definitions of field experts. This question examined how well the indicators of the ECERS-R are representing what the field experts held to be indicative of classroom quality. The first analysis collapsed the agreement categories of "A great deal" and "Somewhat" into one category. The indicators were listed in a column and each expert had a corresponding column with a "1" or a "0" indicating their agreement or non-agreement that that indicator was important for quality assessment. The 1's were summed for each indicator, yielding the number of experts that agreed that each indicator was an important component of quality. The second analysis conducted the same examination as the first but looked at only the highest level of agreement ("A great deal").

Based on the results of these analyses, a new version of the ECERS-R was created that included only those indicators agreed upon by at least 50% of the experts. As in previous analyses, the agreement categories of “A great deal” and “Somewhat” were first collapsed into one category. However, due to lack of variation stemming from most of the experts rating almost all indicators as important to some degree, only those indicators that were deemed to be “A great deal” important were used in the new version of the instrument. Removed indicators are discussed more in the next chapter. This new version of the instrument was referred to as the VU-ECERS (a version created at Vanderbilt University).

Research Question 3: What are the psychometric properties of the ECERS-R and the VU-ECERS?

In order to be used in policies that evaluate the quality of early childhood classrooms, an instrument must meet the basic requirements of reliability and validity in test construction. In addition, the VU-ECERS offers another method for assessing the validity of the current instrument. The ECERS-R and the VU-ECERS were evaluated in the following areas (although all analyses were not possible with both versions of the instrument).

Inter-rater reliability and temporal stability. The inter-rater reliability of the ECERS-R was examined, and the reliability of the ECERS-R and VU-ECERS over time, or their temporal stability, was also analyzed. Quantitative analyses of the inter-rater reliability were not possible due to having to rely on the information provided by each of the sites in the secondary data sets. However, the numbers provided by the sites were compared, when possible, across levels of the ECERS-R to each other and to what the developers report. To assess temporal stability, with those classrooms that were observed twice in the same year, the correlations of item, subscale, and total scores from one time point to the next were calculated. This was done with both versions of the instrument.

Policy implications of observation time. Another area examined was whether the time of observation during the year with the ECERS-R and VU-ECERS affected the score a classroom received. Although one might expect a classroom's score on the instrument to be different at different points in the year due to the teacher getting better over time, the children adapting to the demands of the classroom as the year progresses, and so forth, this is a concern when policy issues are brought into play. Whether scores across the year improve, decline, or exhibit a random pattern across a group of classrooms, the time that an ECERS-R observation is done can mean different reimbursement rates for those classrooms depending on the scores that are reported, especially if the difference in time leads to a change from one side of the cut-off score to the other.

To answer this question, data from the 60 classrooms that were observed twice in the same year (when the lead teacher was the same at both time points) were analyzed. All of these classrooms were given subscale and total instrument scores using the traditional scoring method of the ECERS-R (Method 1), yielding an average total score on a scale from 1 to 7 for each class at each observation point. For each time point, a classroom received a score of 1 if they are above the cut-off point (described in Research Question 1) or a 0 if they are below the cut-off point. Because the nature of the data consisted of a comparison of paired data in which the dependent variable is dichotomous, a McNemar two-tailed nonparametric test was used to examine whether the extent of change in 1's from one time point to another, in either direction, was significant.

Classroom characterization changes from observation time change. As in Research Question 1, classrooms that were observed more than once with the same lead teacher were also analyzed for the effect of time of the year observed on a classroom's standing relative to the ECERS-R quality categories. The VU-ECERS was recoded in the same way as the ECERS-R scores were recoded in Research Question 1.

Data from the 60 classrooms that were observed twice in one year were analyzed. Because the nature of the data consisted of a comparison of paired data in which the dependent variable was categorical, a Wilcoxon matched pairs signed-ranks two-tailed nonparametric test was used to examine whether the extent of change in category from one time point to another, in either direction, was significant.

Observation length and start time as predictors. Another issue surrounding the variable of time is whether or not the length of an observation period and/or the start time of an observation affected the score that a classroom received. For this analysis, the observation length was calculated for each of the observations that included both start time and end time for the observation period. Length and start time were examined as separate predictors of an observation's total score for both the VU-ECERS and ECERS-R instruments.

Internal consistency. The internal consistency of the subscales of the ECERS and the VU-ECERS was also examined. For each of the two instrument versions, Cronbach's Alpha was calculated for each of the six subscales using all of the items under each of the subscales. The Alpha for standardized items (which have been computed to have equal means and variances) was examined along with the change in Alpha when an item was removed.

Concurrent validity. The above-mentioned analyses have all pertained to the reliability of the instrument in different ways. The validity of the scale was also examined through factor analyses. The concurrent validity of an alternative version of an instrument can be shown if the alternative version demonstrates similar factor structures as the original version, and the alternative version has high correlations of factor scores with those of the original scale. For both instrument versions, exploratory factor analyses using varimax rotation were conducted at the item level. First, any items that had missing data for more than 10% of the cases or those items whose distributions were skewed (> +/- 2.0) were removed from the analyses. Removed items are discussed more in-depth in the next chapter. Second, inter-item correlations for each scoring method were examined to check for items with odd patterns of relationships with other items. As in previous studies using data reduction techniques with ECERS-R data (Cassidy et al., 2005), items that loaded on a factor with less than a .4000 factor score were suppressed in the rotated component matrix. The results of the factor analyses were looked at in terms of which items loaded on which factors. In addition, factor scores for classrooms were correlated to yield coefficients pertaining to the concurrent validity of the VU-ECERS, specifically.

Research Question 4: What are the policy implications associated with using the VU-ECERS?

This question looked at how the VU-ECERS changes a classroom's relative ranking from when that classroom was scored with the ECERS-R using the traditional scoring method. First, the VU-ECERS was scored using the traditional method of scoring (only the indicators that were agreed upon by at least 50% of the experts were selected from each classroom's data and included in scoring). The results of the VU-ECERS were compared to ECERS-R scores using the same methods as were used in Research Question 1. Using the methodology described under Research Question 1, this question examined a classroom's degree of change in standing when the content of the instrument is altered.

CHAPTER IV

RESULTS

Initial Analyses

Descriptive Results

Table 6 displays descriptive information about the scores obtained from different scoring methods used on the ECERS-R data for the three secondary data sets (with the Early Math dataset descriptives separate for each state). Because the PCER Tennessee data were not scored at the indicator level, those classrooms were not included in analyses with alternative scoring methods and are only listed in the table under the traditional method. Most classrooms from all three data sets were in the minimal to good range according to ECERS-R scoring divisions. There are small differences in mean values related to the inclusion or exclusion of NA items with each of the summative methods, but the relative ranking of data sets by their mean quality values remains the same across scoring methods.

Table 7 displays the descriptive information for each of the first 6 subscales of the ECERS-R across scoring methods. While the traditional scoring method is not dependent on the number of possible indicators because it is calculated with an average score across items, subscale scores derived using the summative method are independent on the possible number of indicators scored. In order to accurately compare subscale scores across scoring methods, subscale statistics for the summative methods were divided by the number of possible indicators for each subscale, revealing a mean proportion of total possible indicators that were positively

Table 6

Total ECERS-R Quality Scores from Different Scoring Methods by Data Sets

Source	N	Mean	SD	Minimum	Maximum
Traditional Method					
Early Math TN	74	3.81	.94	1.67	5.70
Early Math CA	48	4.75	.91	2.71	6.38
QRS MO	161	4.63	1.28	1.51	6.82
PCER TN	42	3.95	1.08	1.76	6.47
TOTAL	325	4.37	1.19	1.51	6.82
Summative-no stop Method with NAs					
Early Math TN	74	278.22	36.41	159.00	328.00
Early Math CA	48	307.33	28.51	231.00	369.00
QRS MO	146	307.45	51.04	123.00	385.00
TOTAL	268	299.35	45.73	123.00	385.00
Summative-no stop Method, no NAs					
Early Math TN	74	257.30	36.27	147.00	314.00
Early Math CA	48	285.73	24.98	228.00	330.00
QRS MO	146	282.01	45.59	113.00	341.00
TOTAL	268	275.85	41.61	113.00	341.00
Summative-stop Method with NAs					
Early Math TN	74	218.34	45.86	100.00	292.00
Early Math CA	48	260.08	46.29	158.00	355.00
QRS MO	159	256.69	67.80	71.00	378.00
TOTAL	281	247.17	61.66	71.00	378.00
Summative-stop Method, no NAs					
Early Math TN	74	203.91	44.63	87.00	284.00
Early Math CA	48	241.38	42.66	158.00	315.00
QRS MO	159	236.68	60.49	68.00	337.00
TOTAL	281	228.85	55.82	68.00	337.00

Note. The maximum score possible with the traditional method is 7.00; it is 397 with the summative methods including NA's; it is 347 with the summative methods excluding NA's.

scored for each subscale (as well as standard deviation, minimum, and maximum proportions). With the traditional and summative-stop methods, the Personal Care Routines and the Activities subscales had the lowest mean scores out of the six subscales used. However, with the summative-no stop scoring methods, this was not the case, although Activities remained one of the subscales with comparatively lower means than most of the other subscales. Across scoring methods, the subscale with the highest mean was the Interactions subscale.

Table 7

ECERS-R Subscale Scores Across Scoring Methods

Source	N	Mean	SD	Minimum	Maximum
Traditional Method					
Space and Furnishings	325	4.43	1.13	1.63	7.00
Personal Care Routines	325	3.66	1.56	1.00	7.00
Language-Reasoning	325	4.65	1.37	1.00	7.00
Activities	325	3.86	1.20	1.00	7.00
Interaction	325	4.92	1.83	1.00	7.00
Program Structure	325	4.72	1.73	1.00	7.00
Summative-no stop Method with N/As					
Space and Furnishings	268	0.77	0.10	0.34	0.98
Personal Care Routines	268	0.76	0.11	0.39	0.99
Language-Reasoning	268	0.80	0.15	0.15	1.00
Activities	268	0.70	0.16	0.08	0.97
Interaction	268	0.84	0.17	0.23	1.00
Program Structure	268	0.69	0.20	0.09	1.00
Summative-no stop Method, no N/As					
Space and Furnishings	268	0.81	0.11	0.35	1.00
Personal Care Routines	268	0.83	0.11	0.41	0.98
Language-Reasoning	268	0.80	0.15	0.15	1.00
Activities	268	0.72	0.17	0.06	1.01
Interaction	268	0.84	0.17	0.23	1.00
Program Structure	268	0.83	0.18	0.13	1.00
Summative-stop Method with N/As					
Space and Furnishings	281	0.64	0.15	0.23	0.98
Personal Care Routines	281	0.55	0.20	0.13	0.96
Language-Reasoning	281	0.72	0.19	0.08	1.00
Activities	281	0.57	0.18	0.04	0.96
Interaction	281	0.74	0.25	0.09	1.00
Program Structure	281	0.60	0.25	0.00	1.00
Summative-stop Method, no NAs					
Space and Furnishings	281	0.68	0.15	0.25	1.00
Personal Care Routines	281	0.61	0.22	0.14	0.98
Language-Reasoning	281	0.72	0.19	0.10	1.00
Activities	281	0.58	0.19	0.03	1.01
Interaction	281	0.74	0.25	0.09	1.00
Program Structure	281	0.71	0.27	0.00	1.00

Note. Each subscale differs in the total number of indicators possible. Space and Furnishings has 82 indicators (including 5 N/A’s), Personal Care Routines has 77 indicators (including 18 N/A’s), Language-Reasoning has 39 indicators (including 0 N/A’s), Activities has 101 indicators (including 13 N/A’s), Interaction has 53 indicators (including 0 N/A’s), and Program Structure has 45 indicators (including 14 N/A’s).

Research Questions

Research Question 1: To what extent are the results of the ECERS-R affected by the scoring conventions that are currently used by the developers and the two alternative scoring methods used by this researcher?

Scoring method comparison. The relative ranking of classrooms using one scoring method was correlated with the relative ranking of classrooms using a different scoring method. Table 8 displays the results of the relative ranking method comparison. All of the correlations were significant and very high.

Though the correlations for all comparisons were highly significant, there was substantial movement in ranking for many classrooms when the scoring method was altered. To analyze the ranking change for each comparison, the ranking of the second method listed was subtracted from the ranking of the first method. For example, in the first comparison of the traditional method compared to the summative-no stop method including NA items, the ranking of a classroom using the summative-no stop method was subtracted from that same classroom's ranking using the traditional method.

In each of the comparisons, less than 10% of the classrooms did not change rank when the scoring method was altered. In each comparison, the change associated with most classrooms' rank was relatively small (10 slots or less), although some comparisons saw up to 38 classrooms jumping 30 slots or more when the scoring method changed. The high correlations are possible even with the extent of movement in ranking because the majority of classrooms that did change rank moved relatively few slots. In each of the comparisons, over 70% of the classrooms that changed ranking moved less than 30 slots up or down. This percentage was

Table 8

Comparison of Scoring Methods on Classroom Ranking Relative to Each Other with ECERS-R Data

			Classes	Classes Moving Up in Rank				Classes Moving Down in Rank			
Source	N	Pearson *r*	Not Moving	N	Range	10 or >	30 or <	N	Range	10 or >	30 or <
Traditional v. summative-no stop with NAs included	268	.939	7	126	1-81 slots	41	38	135	1-74 slots	51	34
Traditional v. summative-no stop, no NAs included	268	.957	9	128	1-81 slots	50	31	131	1-80 slots	53	25
Traditional v. summative-stop with NAs included	281	.984	9	141	1-44 slots	79	4	131	1-51 slots	79	12
Traditional v. summative-stop, no NAs included	281	.986	14	132	1-36 slots	79	6	135	1-39 slots	81	10
Summative-no stop v. summative-stop, with NAs included	268	.962	10	136	1-63 slots	60	24	122	1-71 slots	50	26
Summative-no stop v. Summative-stop, no NAs included	268	.966	8	138	1-64 slots	69	22	122	1-62 slots	44	19

Note. All correlations are significant at 0.001 level (2-tailed).

highest for the comparisons of traditional and summative-stop with and without NA items (94% of classrooms in each of these comparisons moved less than 30 slots up or down), which are also the comparisons for which the highest correlation between rankings was seen.

Because results of the comparisons for scoring methods with NA items included or without NA items included were almost identical, the summative methods that excluded the NA items were used for the remainder of the analyses in this research question.

Policy implications of scoring method. A second analysis concerned the extent of change in a classroom's location above or below the quality funding cut-score (using the state of Tennessee cut-score of 4.0) when the scoring method was altered. Table 9 shows the amount of change around that cut-score when different scoring methods were used. The McNemar two-tailed nonparametric test revealed that a significant number of classrooms changed their location relative to the funding cut-score in each scoring method comparison. In the comparison of traditional and summative-no stop methods, while no classrooms that were above the cut-score in the first method moved below the cut-score with the second, over 25% of the classrooms moved

Table 9

Classrooms Changing Location Relative to the Quality Funding Cut-Score Using the ECERS-R

Source	Classrooms Below the Cut-Score With Both Methods	Classrooms Above the Cut-Score With Both Methods	Classrooms Changing from Below to Above	Classrooms Changing from Above to Below
Traditional v. Summative-no stop	28	164	76	0
Traditional v. Summative-stop	112	147	0	22
Summative-no stop v. Summative-stop	28	143	0	97

from below the cut-score with the traditional method to above the cut-score with the summative-no stop method. This trend was reversed in all other method comparisons in the table where between 7.8% and 38% of classrooms moved from above the cut-score with the first method to below the cut-score with the second method.

Classroom characterization by alternate scoring methods. A third analysis looked at the extent to which classrooms would change ECERS-R total quality categories when a different scoring method was used, and what direction those changes would be in. The ECERS-R has four categories of quality represented in the scoring scale. A score of "1" is considered Inadequate, a score of "3" is considered Minimal, a score of "5" is considered Good, and a score of "7" is considered Excellent. Table 10 shows the number of classrooms that moved to another category when a different scoring method was applied to their ECERS-R data, as well as the direction of the change. The Wilcoxon matched pairs signed-ranks two-tailed nonparametric tests revealed

Table 10

Classrooms Changing ECERS-R Quality Categories with ECERS-R Data

Source	Classrooms Not Changing	Classrooms Moving Down	Classrooms Moving Up	Z
Traditional v. Summative-no stop	166	0	102	-10.10
Traditional v. Summative-stop	247	34	0	-5.38
Summative-no stop v. Summative-stop	133	135	0	-11.62

Note. All statistics were significant at the .001 level (2-tailed).

significant category change with all scoring method comparisons. For every classroom that changed categories, whether up or down, no classroom moved more than one category in either direction.

In the comparison relative to the quality cut-score in Table 9, the pattern of movement showed no classrooms moving down in the comparison of the traditional and summative-no stop methods, and no classrooms moving up in any of the other comparisons. The same pattern of movement existed in the analysis of ECERS-R category movement. In each comparison, regarding the classrooms that changed categories when the scoring method was altered, movement was seen across all of the categories. The majority of movement up or down across scoring methods, however, involved movement from Minimal to Good or vice-versa.

Research Question 1 Summary

This question examined how the results of the ECERS-R were affected by using different scoring methods on the same data. The total scores are stable across scoring methods; the correlations between scores using different scoring methods are very high. However, when those scores are used for comparison purposes (i.e. comparing classrooms to each other, comparing classrooms to a funding cut-score, or comparing classrooms to ECERS-R quality categories), they do not behave in the same way across scoring methods. Concerning the ranking of classrooms relative to each other, only a small percentage of classrooms remained in the same ranked standing when the scoring method was changed. Although the majority of classrooms that moved did not move very far, the implications of that movement are more evident when the potential impact of alternative scoring methods on funding is considered.

When examining the effect of alternate scoring methods on a classrooms standing relative to a quality cut-score that is used in Tennessee, as high as 28% of classrooms that would've been under that cut-score when the traditional method was used would have been eligible (aside from the other variables that are involved in determining funding) for funding when the summative-no stop method was used. When the summative-stop scoring method was used, no classroom originally scored using the traditional or the summative-no stop method moved from below the cut-score to above the cut-score.

Research Question 2: To what extent do field experts agree among themselves and with the instrument's developers on the organization and content of the ECERS-R?

Agreement on organization. To answer the question about how much experts agreed with ECERS-R developers about the organization of indicators into subscales, first the agreement categories of "A great deal" and "Somewhat" were collapsed into one variable representing subscale placement agreement for each respondent. Agreement between experts and developers for each indicator was determined by comparing the subscale that each respondent placed an indicator in and the subscale that the ECERS-R developers placed that indicator in. The percentage of all 397 ECERS-R indicators from the first six subscales for which both the ECERS-R developers and the experts agreed about subscale placement was calculated for each expert. Table 11 displays the percent agreement for each expert and for the expert panel as a whole at both the subscale and total score levels. The total percentage for each expert represents the total percentage of indicators for which subscale placement was agreed upon by the expert and the ECERS-R developers. The total percentage for the panel for each subscale represents the mean percentage of indicators under each subscale on which the experts agreed with the

Table 11

Subscale Agreement Between ECERS-R Developers and Experts, Agreement Collapsed

	Percentage of Indicators Agreed Upon Under Each Subscale						
Source	Space & Furnishings	Personal Care Rtnes	Language-Reasoning	Activities	Interaction	Program Structure	Total
Expert 1	93.90	88.31	94.87	93.07	100.00	80.00	91.94
Expert 2	93.90	72.73	92.31	83.17	94.34	62.22	83.38
Expert 3	97.56	96.10	100.00	96.04	98.11	86.67	95.97
Expert 4	95.12	87.01	97.44	87.13	96.23	100.00	92.44
Expert 5	92.68	66.23	71.79	81.19	90.57	75.56	80.35
Expert 6	93.90	74.03	53.85	72.28	75.47	91.11	77.83
Expert 7	82.93	62.34	71.79	65.35	69.81	28.89	65.49
Expert 8	95.12	80.52	87.18	74.26	92.45	44.44	80.10
Expert 9	89.02	37.66	82.05	81.19	92.45	88.89	76.83
Expert 10	100.00	77.92	100.00	98.02	100.00	93.33	94.46
Expert 11	82.93	66.23	89.74	73.27	66.04	73.33	74.56
Expert 12	90.24	41.56	56.41	45.54	86.79	46.67	60.71
Expert 13	82.93	48.05	89.74	42.57	71.70	57.78	62.22
Expert 14	87.80	51.95	48.72	85.15	86.79	73.33	74.56
Expert 15	79.27	83.12	82.05	97.03	83.02	62.22	83.38
Expert16	97.56	76.62	92.31	91.09	90.57	24.44	82.12
Total for Panel	90.93	69.40	81.89	79.15	87.15	68.06	79.77

developers about placement. For each expert, the total percentage of indicators for which there was agreement between the respondent and the developers ranged from 60.71% to 95.97% with an average total for the panel of 79.77% (SD=10.83%), indicating a fairly high average amount of agreement between respondents and developers concerning the organization of the ECERS-R when the two agreement categories were collapsed. However, though overall agreement was high, the level of agreement differed at the subscale level. For Subscales 1, 3, and 5 (Space & Furnishings, Language-Reasoning, and Interaction), experts and developers agreed that, on average, over 80% of the indicators in that subscale belonged there. For Subscales 2 and 6 (Personal Care Routines and Program Structure), experts and developers agreed that, on average, less than 70% of the indicators in that subscale belonged there.

The percentage of indicators that placement was agreed upon by all experts on the panel was also calculated at both the subscale and total levels (with the two agreement categories collapsed). For a given domain (subscale), this analysis looked at the percent of all 397 indicators that every expert put under the same domain (regardless of where the ECERS-R developers had placed that indicator). Table 12 displays the percentage of agreement among experts. Though there were two subscales for which all experts agreed among themselves on the placement of at least 50% of the indicators (Space and Furnishings and Interaction), there were also two subscales for which all experts agreed on the placement of fewer than 10% of the indicators (Language-Reasoning and Program Structure). For the overall scale, the percentage of indicators for which there was agreement on placement among all of the experts was fairly low (approximately ¼ of the total indicators on the scale).

Table 12

Organization Agreement Among Expert Panel Members, Agreement Collapsed

Source	Percentage of Indicators Agreed Upon
Space and Furnishings subscale	50.00
Personal Care Routines subscale	15.58
Language-Reasoning subscale	5.13
Activities subscale	13.86
Interaction subscale	56.60
Program Structure subscale	6.67
Total Scale	25.69

In addition to examining the percentage of *indicators* on which placement was agreed by both experts and developers, the researcher also calculated the percentage of the group of *experts* who agreed with the developers about the subscale placement of each indicator. Across all 397 indicators, 94% of the experts agreed with the ECERS-R developers about the placement of 95 indicators, or 23.93% of the indicators. At least one expert agreed with the developers about the subscale placement for every indicator. At least half of the experts agreed with the developers' placement of 340 indicators, or 85.64% of the indicators. The mean percent of experts agreeing with ECERS-R developers about the subscale placement of indicators was 74.64% (SD=19.99%). It is important to note that the two agreement categories of "A great deal" and "Somewhat" were still collapsed in these analyses.

It is also important to note that experts may have placed any indicator under more than one subscale, although multiple subscale placement of indicators was not considered in these analyses. Only one respondent did not put any indicators into more than one subscale. For the analysis of the percentage of indicators for which there was agreement between each respondent and the developers, any indicator that an expert placed under the same subscale as the developers was counted as agreed upon, regardless of how many other subscales that expert placed each indicator. For the analysis of the percentage of indicators that placement was agreed upon by all experts on the panel, an indicator that all 16 experts placed under the same subscale was counted as agreed upon by the panel, regardless of whether some experts may have also placed that indicator under other subscales. For the analysis of the percentage of experts who agreed with the developers about the subscale placement of each indicator, an expert was counted as agreeing with the developers upon placement for an indicator if the expert had placed that indicator under the same subscale as the developers did, regardless of whether the expert had also placed that indicator under other subscales.

After the previous analyses were conducted with the two agreement categories collapsed, the analyses were re-run without the categories collapsed, using only the anchor of "A great deal" of agreement to represent an expert's placement of an indicator under a given subscale. For example, for Indicator 17.3.1, "Staff sometimes talk about logical relationships or concepts," 15 of the 16 experts agreed with the developers that the indicator belonged under the Language-Reasoning subscale. However, of those 15 experts, 2 indicated that they agreed with the placement only "Somewhat" instead of "A great deal." When agreement categories were not collapsed, only 13 experts would be counted as agreeing with ECERS-R developers on the placement of Indicator 17.3.1, while when agreement categories were collapsed, 15 experts would be counted as agreeing with developers.

Table 13 displays the percent agreement for each expert and for the expert panel as a whole at both the subscale and total score levels. The same general pattern of subscale agreement can be seen in Table 13. The dimensions with the highest percentage of indicators on which placement was agreed upon between respondents and developers were Subscale 1 and Subscale 5 (Space and Furnishings and Interaction), with the lowest percentage of agreement seen in Subscale 6 (Program Structure). For the total scale, agreement between experts and developers decreased from almost 80% of indicators (with agreement collapsed) to 57% (with only the highest level of agreement).

Table 13

Subscale Agreement Between ECERS-R Developers and Experts, Agreement Not Collapsed

	Percentage of Indicators Agreed Upon						
Source	Subscale1	Subscale2	Subscale3	Subscale4	Subscale5	Subscale6	Total
Expert 1	90.24	74.03	82.05	71.29	79.25	71.11	77.83
Expert 2	68.29	53.25	61.54	31.68	81.13	11.11	50.63
Expert 3	87.80	72.73	84.62	41.58	83.02	55.56	68.51
Expert 4	92.68	64.94	64.10	31.68	75.47	62.22	63.22
Expert 5	90.24	58.44	61.54	44.55	83.02	68.89	66.25
Expert 6	57.32	44.16	25.64	22.77	45.28	37.78	39.04
Expert 7	18.29	24.68	10.26	8.91	43.40	22.22	20.15
Expert 8	90.24	62.34	87.18	65.35	81.13	11.11	68.01
Expert 9	89.02	37.66	82.05	81.19	92.45	88.89	76.83
Expert 10	65.85	53.25	46.15	61.39	71.70	24.44	56.42
Expert 11	51.22	53.25	76.92	36.63	41.51	42.22	48.11
Expert 12	71.95	41.56	51.28	38.61	86.79	44.44	54.41
Expert 13	69.51	45.45	89.74	26.73	67.92	53.33	53.90
Expert 14	50.00	36.36	41.03	47.52	62.26	37.78	46.10
Expert 15	59.76	45.45	48.72	51.49	64.15	35.56	51.64
Expert 16	86.59	53.25	82.05	53.47	79.25	17.78	62.47
Total for Panel	71.19	51.30	62.18	44.68	71.11	42.78	56.47

The percentage of indicators that placement was agreed upon by all experts on the panel was also calculated at both the subscale and total levels without collapsed agreement categories. None of the 397 indicators had full agreement by all of the experts on their subscale placement, indicating that while respondents generally agreed with developers on the placement of roughly a quarter of the indicators, respondents' degree of agreement on each of those indicators varied.

The researcher also calculated the percent of experts that agreed with the developers about the subscale placement of each indicator without collapsed agreement. Across all 397 indicators, 94% of the experts agreed with the ECERS-R developers about the placement of 14 indicators, or 3.53% of the indicators. For 8 of the indicators, or 2.02% of the indicators, none of the experts agreed with the developers about the subscale placement. At least half of the experts agreed with the developers' placement of 240 indicators, or 60.45% of the indicators. The mean percent of experts agreeing with ECERS-R developers about the subscale placement of indicators was 52.57% (SD=25.56%). It is important to note that this analysis did not include indicators that experts agreed with "Somewhat," and that experts may have placed any indicator under more than one subscale, although that was not considered in this analysis.

Agreement on content. Another issue regarding the Indicator Survey was the extent to which field experts agreed that the ECERS-R indictors were important components of quality. This group of analyses was independent of the subscale under which experts placed indicators. Table 14 displays the number of indicators that were rated with each of the survey options by each of the experts and the panel as a whole. For the panel as a whole, the majority of ECERS-R indicators were considered to be very important. Only a small percentage of indicators were rated as having no importance at all to the panel's definition of quality. There was substantial range, though, among the experts as to their importance ratings for each of the indicators. For example, the percent of indicators rated as having no importance ranged from 0% to 44% across experts.

Table 14

Expert Rating of the Importance of ECERS-R Indicators to Their Definitions of Quality

	Number of Indicators Rated as Having		
Source	A great deal of importance	Some importance	No importance
Expert 1	191 (48.35%)	166 (42.03%)	38 (9.62%)
Expert 2	172 (43.32%)	186 (46.85%)	39 (9.82%)
Expert 3	344 (86.87%)	50 (12.63%)	2 (0.51%)
Expert 4	321 (81.06%)	75 (18.94%)	0 (0.00%)
Expert 5	280 (70.89%)	114 (28.86%)	1 (0.25%)
Expert 6	234 (58.94%)	157 (39.55%)	6 (1.51%)
Expert 7	246 (62.28%)	149 (37.72%)	0 (0.00%)
Expert 8	236 (59.90%)	130 (32.99%)	28 (7.11%)
Expert 9	358 (90.40%)	29 (7.32%)	9 (2.27%)
Expert 10	260 (65.49%)	126 (31.74%)	11 (2.77%)
Expert 11	306 (77.47%)	84 (21.27%)	5 (1.27%)
Expert 12	340 (86.29%)	49 (12.44%)	5 (1.27%)
Expert 13	262 (66.67%)	96 (24.43%)	35 (8.91%)
Expert 14	232 (59.18%)	117 (29.85%)	43 (10.97%)
Expert 15	208 (53.20%)	150 (38.36%)	33 (8.44%)
Expert 16	152 (38.38%)	200 (50.51%)	44 (11.11%)
Total for Panel	4142 (65.55%)	1878 (29.72%)	299 (4.73%)

The number of experts that rated each indicator as important to their personal definitions of quality was also calculated. This analysis was first done with the two ratings of “A great deal” and “Somewhat” collapsed into one agreement category. With these ratings collapsed, there were 216 indicators, or 54.41% of the ECERS-R indicators from the first six subscales, that all experts agreed were important to quality. Each indicator was listed by at least one expert as important to that person’s individual definition of quality. The range of the number of experts rating each indicator as important was 6 to 16. Across all 397 indicators, the average number of experts who said that each indicator was important to their definition of quality was 15.16 (SD=1.38).

The same analysis was conducted without the collapsed agreement categories. In this analysis, only those indicators that an expert rated with “A great deal” of agreement were counted. There was only one indicator that none of the experts rated as important. This indicator was “Staff are actively involved in use of TV, video, or computer.” There were 25 indicators that were rated as important by all 16 experts (6.30% of the rated indicators). The range of the number of experts rating each indicator as important was 0 to 16 with a mean of 10.43 (SD=3.94). The number of indicators that at least 8 of the 16 experts rated as important to quality was 300, or 75.57% of the rated indicators.

Indicator importance was also examined at the subscale level to get information about the percent of each subscale’s indicators were agreed upon. Indicators were grouped according to the subscales that ECERS-R developers had placed them in, and the number of experts rating each of those indicators as being of great importance to quality was calculated. Table 15 displays the percent of each subscale’s indicators that at least 50% of the experts rated as important to their definition of quality.

Table 15

Percent of Indicators that Experts Rated as Important for Each Subscale

Subscale	Percentage of Indicators Rated as Important
Space and Furnishings subscale	54.88%
Personal Care Routines subscale	85.71%
Language-Reasoning subscale	100.00%
Activities subscale	54.46%
Interaction subscale	96.23%
Program Structure subscale	97.78%

For subscale 3, Language-Reasoning, at least 50% of the respondents rated every indicator under that ECERS-R subscale as important to quality. However, for the first and fifth subscales, Space and Furnishings and Interaction, approximately half of the indicators were rated as unimportant by at least half of the respondents.

With the data obtained from the Indicator Survey analysis, a new version of the ECERS-R, the VU-ECERS, was created using only those indicators that at least 50% of the experts agreed were of great importance to quality. The VU-ECERS was comprised of 300 indicators, with 97 indicators from the original ECERS-R removed. Out of all 97 removed indicators, 11 indicators were NA indicators on the ECERS-R. For one item, Furnishings for relaxation and comfort, all indicators were removed, in effect removing the entire item. The number of removed indicators is listed in Table 16 by subscale and item under each of the ECERS-R quality categories. A complete list of removed indicators is provided in Appendix B. No indicators were removed from Subscale 3. Subscale 4, Activities, had the most indicators removed from the most items. The last subscale, Program Structure, had the least indicators removed from the least items.

Table 16

Number of Indicators Removed from ECERS-R for VU-ECERS

Source	Inadequate Indicators Removed	Mimimal Indicators Removed	Good Indicators Removed	Excellent Indicators Removed	Total Indicators Removed
Subscale 1	5	4	15	13	37
Indoor Space	0	0	1	1	2
Furniture for routine care, play, and learning	0	0	0	2	2
Furniture for relaxation	2	2	3	2	9
Room arrangement for play	0	0	3	1	4
Space for privacy	1	1	2	2	6
Child-related display	1	1	2	2	6
Space for gross motor	0	0	3	2	5
Gross motor equipment	1	0	1	1	3
Subscale 2	1	1	6*	3*	11*
Greeting/departing	1	1	1	0	3
Meals/snacks	0	0	1	1	2
Nap/rest	0	0	3**	1**	4**
Toileting/diapering	0	0	0	1	1
Health Practices	0	0	1	0	1
Subscale 4	2	10	19*	15*	46*
Fine motor	0	1	2	1	4
Art	0	0	0	3*	3*
Music/movement	0	3	2	3	8
Blocks	0	1	4	3	8
Sand/water	2	2	3	0	7
Dramatic play	0	1	2	3	6
Nature/science	0	1	2	0	3
Use of TV, video, and/or computers	0	1**	3**	2**	6**
Promoting acceptance of diversity	0	0	1	0	1
Subscale 5	0	1	0	1	2
Supervision of gross motor activities	0	0	0	1	1
General supervision of children	0	1	0	0	1
Subscale 6	0	1	0	0	1
Schedule	0	1	0	0	1
Total	8	17*	40*	32*	97*

**Note.* Some indicators in this column were possible NA indicators.

***Note.* All indicators for this item were possible NA indicators.

In order for the VU-ECERS to be scored with the traditional ECERS-R scoring method, some items had to be removed. An item was removed if, after deleting the indicators that were not deemed important by at least half of the expert panel, that item did not have at least one indicator under each of the ECERS-R anchors. If items that had no indicators under the “Good” anchor, for example, were included, the possible variation in scores for that item would be decreased. Using the traditional scoring method, it would not be possible for a classroom to score a 4, 5, or 6. Therefore, 11 entire items were removed from the VU-ECERS data. Those items included 2, 4, 5, 6, 7, 11, 20, 21, 22, 23, and 27. Item 3 had already been removed because all of its indicators had been deleted. Data from the PCER data set was not able to be scored on the VU-ECERS because that data set did not include scores at the indicator level, and so indicators that were not agreed on by the expert panel could not be removed from the scores of those classrooms. Table 17 displays the average VU-ECERS scores for classrooms by data set.

Table 17

Total VU-ECERS Quality Scores by Data Sets

Source	N	Mean	SD	Minimum	Maximum
Early Math TN	74	3.89	0.87	1.63	5.60
Early Math CA	48	4.94	0.97	2.85	6.52
QRS MO	150	4.77	1.32	1.52	7.00
TOTAL	272	4.56	1.22	1.52	7.00

In Table 18, the average subscale scores on the VU-ECERS are displayed. As was true of the ECERS-R subscale scores using the traditional scoring method, the Personal Care

Routines subscale had the lowest mean score, while the highest mean score was seen for the Interactions subscale.

Table 18

VU-ECERS Subscale Scores

Source	N	Mean	SD	Minimum	Maximum
Space and Furnishings	272	4.56	1.72	1.00	7.00
Personal Care Routines	272	3.92	1.61	1.20	7.00
Language-Reasoning	272	4.71	1.36	1.00	7.00
Activities	272	4.26	1.23	1.00	7.00
Interaction	272	5.04	1.83	1.00	7.00
Program Structure	272	4.86	1.75	1.00	7.00

Research Question 2 Summary

This question examined the extent of agreement between field experts and ECERS-R developers on the organization of the indicators and the indicators' importance to the personal definitions of quality that the experts held. In terms of agreement between experts and ECERS-R developers on the organization of the instrument, there was a high degree of agreement when the two agreement categories were collapsed. When just the highest degree of agreement was examined, experts agreed with developers on the placement of less than 60% of the indicators, with the percentage of indicators agreed upon ranging from 20% to 77% across experts. There was differential agreement at the subscale level. The highest agreement was seen for the placement of indicators in the Space and Furnishings subscale, with the least agreement concerning the subscales of Personal Care Routines and Program Structure. There was also a good deal of disagreement among experts, especially for indicators under the Personal Care

Routines and Program Structure subscales. More disagreement on indicator placement was found among experts than between experts and ECERS-R developers.

In terms of agreement between experts and ECERS-R developers on the content of the instrument, the panel agreed that most indicators were important to quality at some level, but there was a lot of disagreement among experts about how important the indicators were ("A great deal" or "Somewhat"). At the subscale level, Space and Furnishings and Activities had the least number of indicators rated as having a great deal of importance to quality, while the Language-Reasoning, Interaction, and Program Structure subscales had over 95% of their indicators agreed on by experts at the highest level of importance. Most of the indicators that all experts did not agree were of great importance to quality were at the higher anchors of the ECERS-R scale, as opposed to those indicators that represent Inadequate or Minimal quality.

Research Question 3: What are the psychometric properties of the ECERS-R and the VU_ECERS?

ECERS-R Inter-rater reliability and temporal stability. In order to be used in policies that evaluate the quality of early childhood classrooms, an instrument must meet the basic requirements of reliability and validity in test construction. The inter-rater reliability of observers using the ECERS-R in the secondary data sets was compared using the three scoring methods. Data on the reliability of observers from the RTI study were not provided to this researcher. The Missouri project did not collect data on the reliability of observers at the indicator level.

Table 19 shows the inter-rater reliabilities of observers at different levels of the ECERS-R, as well as the reliability of observers reported by the developers of the ECERS-R. Those levels correspond to the different methods of scoring. The traditional method would use reliabilities at the item level, while both of the summative methods would use reliability at the indicator level. Reliability among observers was higher in all three sources of secondary data than what was reported by the instrument's developers. However, in all sources in the table, reliability was highest at the indicator level.

Table 19

Inter-rater Reliability of ECERS-R Observers

Source	Observer N	Item-Level Exact Agreement	Item-Level Agreement Within One	Indicator-Level Exact Agreement
CA Early Math Project	5	63%	78%	87%
TN Early Math Project	6	61%	78%	88%
MO QRS Pilot Study	18	85%	90%	(not reported)
ECERS-R Developers	2	48%	71%	86%

Note. Information from the ECERS-R developers on inter-rater reliability was reported in the instrument, itself. The developers, like the QRS Pilot Study, reported reliability based on all 7 subscales of the instrument as opposed to the 6 subscales used in the other two studies in the table.

In addition to the reliability of the observers, the reliability of the instrument over time was also examined. Instruments that are used to evaluate programs should be stable across time when classroom circumstances have not changed substantially. Table 20 displays the correlations between ECERS-R scores from classrooms that were observed twice in the same academic year when the lead teacher remained the same. Correlations are given at the total and subscale levels. All correlations were significant. However, there is the suggestion from one of the data sets that ECERS-R scores may not be reliable over a few months time. The same temporal stability analysis was conducted for the traditional method with Missouri classrooms and Tennessee (PCER) classrooms separately. For the 39 Missouri classrooms, all correlations were still highly significant at the total and subscale levels. For the 21 Tennessee classrooms, the correlation between the total scores was not significant (*r*=.384, p>.05). In addition, only two of the subscales, Language-Reasoning and Activities, had significant correlations at the .05 level from one time point to the next (*r*=.534 and *r=.536*, respectively). Although those correlations were statistically significant, the meaning behind the correlations is that only 28% of the variation in ECERS-R scores at the second time point could be explained by scores at the first time point.

Table 20

Temporal Stability of Total and Subscale ECERS-R Scores

Source	N of Classrooms	Pearson *r*	p-value
Space and Furnishings	60	.570	.000
Personal Care Routines	60	.649	.000
Language-Reasoning	60	.590	.000
Activities	60	.553	.000
Interaction	60	.391	.002
Program Structure	60	.623	.000
Total Score	60	.656	.000

In addition to looking at the temporal stability of ECERS-R scores at the total and subscale levels, the temporal stability of scores at the item level was also examined. Table 21 shows the correlations between two time points for ECERS-R item scores in classrooms that were observed twice in the same year. Nonsignificant correlations are bolded.

Table 21

Temporal Stability of Item-Level ECERS-R Scores (N = 60)

Item	*r*
Indoor Space	0.458
Furniture for care, play, learning	**0.138**
Furniture for relaxation	0.437
Room arrangement	0.556
Child-related display	0.546
Space for privacy	0.298
Space for gross motor	0.516
Gross motor equipment	0.422
Greeting/departing	0.314
Meals/snacks	0.551
Nap/rest	0.396
Toileting/diapering	0.530
Health practices	**0.157**
Safety practices	0.549
Books & pictures	0.309
Encouraging children to communicate	0.331
Using language to develop reasoning	0.555
Informal use of language	0.536
Fine motor	0.372
Art	0.280
Music/movement	0.355
Blocks	0.528
Sand/water	0.262
Dramatic play	**0.200**
Nature/science	0.618
Math/number	0.300
Use of TV, radio, computers	**0.237**
Promoting acceptance & diversity	0.367
Supervision of gross motor activities	0.320
General supervision of children	**0.246**
Discipline	0.375
Staff-child interactions	0.406
Interactions among children	0.334
Schedule	0.517
Free play	0.459
Group time	0.619
Provision for children with disabilities	**0.200**

Note. Nonsignificant correlations are bolded.

At the item level, 77% of the significant correlations were significant at the .01 level (two-tailed). A lesser portion, 23%, were significant at the .05 level (two-tailed). The significant correlations ranged from .262 to .619. Of those correlations, 61% of them were less than .500, meaning that in over half of the items that were significantly correlated from one time point to the next, only 25% or less of the variation in scores at Time 2 could be predicted by scores at Time 1.

Again, there is the suggestion from one of the data sets that ECERS-R item scores may not be reliable over a few months time. When Tennessee (PCER) and Missouri classrooms were analyzed separately, the pattern of significance across item correlations was different. While 83% of the items still showed significant correlations across time for the Missouri classrooms, only 19% of the items showed such correlations for the Tennessee classrooms.

Policy implications of observation time for ECERS-R scores. Regarding temporal stability, this study also examined changes in a classroom's standing relative to the quality funding cut-score when observed at different times of the academic year. Table 22 displays the results of the McNemar two-tailed nonparametric test that was used to examine whether the extent of change from one side of the cut-score to the other across time was significant.

Table 22

Temporal Stability of Total ECERS-R Scores Relative to Funding Quality Cut-Scores

Scoring Method	Classrooms Below the Cut-Score At Both Times	Classrooms Above the Cut-Score At Both Times	Classrooms Changing from Below to Above	Classrooms Changing from Above to Below
Traditional	14	27	7	12

According to the test, there was not significant change relative to the quality cut-score. However, when Tennessee (PCER) and Missouri classrooms were analyzed separately, there was significant change associated with Tennessee classrooms but not with Missouri classrooms.

Classroom characterization changes from observation time change for ECERS-R scores. Changes in a classroom's standing relative to the ECERS-R quality categories from one time point to the next were examined. Table 23 displays the results of the Wilcoxon matched pairs signed-ranks two-tailed nonparametric test that was used to examine whether the extent of change in categories across time was significant using each scoring method. According to the test, there was not a significant amount of change from one category to another over time.

Table 23

Temporal Stability of Total ECERS-R Scores Relative to ECERS-R Quality Categories

Scoring Method	Classrooms Not Changing	Classrooms Moving Down	Classrooms Moving Up	Z
Traditional	36	9	15	-1.225

However, when Tennessee and Missouri classrooms were analyzed separately, there was significant category change associated with the Tennessee classrooms but not for the Missouri classrooms.

Observation length and start time as predictors of ECERS-R scores. An important consideration when using an observational measure in classrooms is how long the observation needs to be to get a relatively stable estimate of the behavior being assessed. This study looked at how the length of observation time predicts ECERS-R scores. Information on length of observation period was not provided for the data collected by RTI, and 11 classrooms from the other data sets were missing either the beginning or ending time of the observation, so 53 classrooms were omitted from these analyses. Across the 272 remaining classrooms, the mean observation length was 4.14 hours with a standard deviation of 1.37 hours. The minimum observation length was 1.50 hours and the maximum length was 7.17 hours. Table 24 shows the results of the linear regressions run with observation length as the predictor and total and subscale scores as dependent variables.

Table 24

Predicting Scores from Observation Length Using ECERS-R

Source	Standardized β	*t*-value	p-value
Space and Furnishings	-0.206	-3.451	0.001
Personal Care Routines	0.017	0.284	0.777
Language-Reasoning	-0.055	-0.897	0.371
Activities	-0.222	-3.748	0.000
Interaction	-0.137	-2.777	0.006
Program Structure	-0.333	-5.795	0.000
Total Score	-0.201	-3.374	0.001

All of the significant *t*-values were negative, indicating that longer observation periods predicted lower ECERS-R scores. An example of this relationship can be seen in Figure 2, which plots observation length on the x-axis and total ECERS-R scores from the traditional scoring method on the y-axis. Though significant variation in total scores is predicted by observation length across methods, the same is not true at the subscale level. Subscales 2 and 3, Personal Care Routines and Language-Reasoning, are not influenced by observation length.

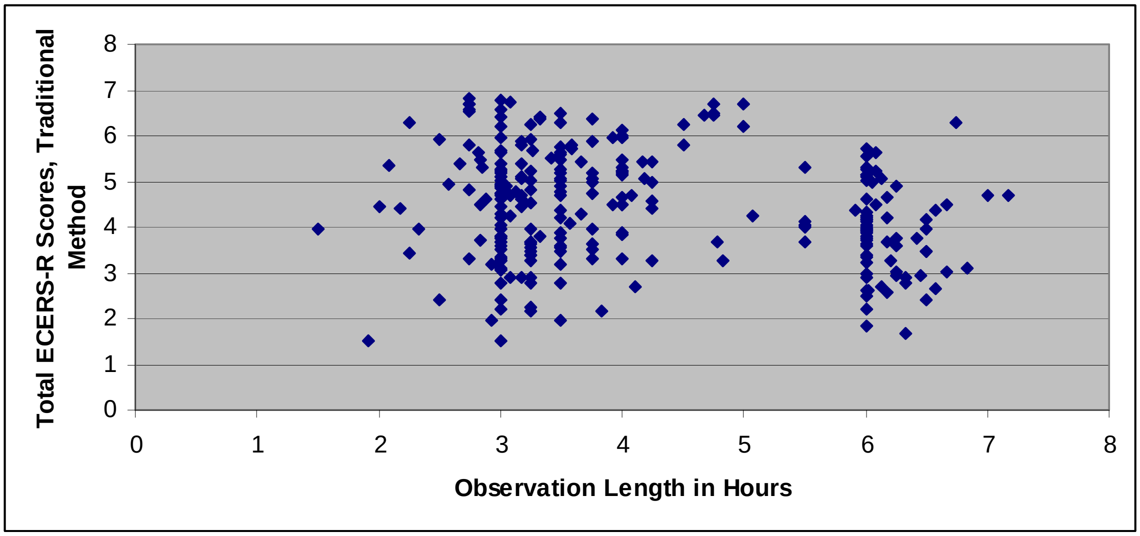


Figure 2. Relationship between length of observation and total ECERS-R scores using the traditional scoring method.

Along with observation length, the start time of the observation was also examined as a predictor of ECERS-R scores. The earliest observation began at 7:00 a.m., while the latest observation began at 12:45 p.m. The average start time was 8:49 a.m. (SD=1.17 hours). A regression analysis was conducted with observation start time as the predictor and total ECERS-R scores as the dependent variable. The analysis revealed that the start time of the observation was not a significant predictor of ECERS-R scores.

Internal consistency of ECERS-R. The internal consistency of the ECERS-R subscales was examined by calculating Cronbach's Alpha for the entire scale and looking at the effects on that Alpha if a subscale was removed. The correlations of subscales to total scores were also examined. Table 25 displays the Alpha, change in Alpha, and subscale to total correlations. The scale showed a high level of internal consistency. The overall alpha was decreased with the removal of every subscale, but it never decreased by more than .04. All subscales had relatively high correlations with the total scale, with correlations ranging from .731 to .848. The highest subscale to total correlation was seen for the Program Structure subscale, and the lowest correlation was seen for the Personal Care Routines subscale. Numbers were similar to what the ECERS-R developers reported in the manual for the same analysis with all seven subscales.

Table 25

Internal Consistency for ECERS-R Subscales with ECERS-R Data

Source	Alpha (based on Standardized items)	Decrease in Alpha if Subscale is Deleted	Subscale to Total Correlation
Total Scale	0.897		
Space and Furnishings		0.031	0.815
Personal Care Routines		0.013	0.731
Language-Reasoning		0.037	0.833
Activities		0.030	0.801
Interaction		0.030	0.836
Program Structure		0.037	0.848

Note. All correlations were significant at the .01 level (two-tailed).

Concurrent validity of ECERS-R. Another important psychometric consideration for an instrument used so widely is the construct validity of the scale. Factor analyses of the items were conducted using a principal component extraction method with varimax rotation. First, inter-subscale correlations were examined (aggregate statistics displayed in Table 26), which is a different analysis than the subscale to total score correlations that were examined in the analysis of the scale’s internal consistency. If the instrument with six subscales is truly a measure of six distinct components of quality, correlations among subscales should not be too low as to indicate that a subscale might measure a construct unrelated to the other subscales but should likewise not be too high as to indicate that less than six distinct components are being assessed. While all inter-subscale correlations were significant, they were not high enough to indicate that subscales were duplicative.

Table 26

Inter-Subscale Correlations for ECERS-R Scores

Scoring Method	N	Mean *r*	Correlation Range
Traditional	325	0.591	.399-.734

Note. All correlations were significant at the .01 level (two-tailed).

Next, the inter-item correlations were examined to check for items that did not correlate with many of the other items or items that correlated too highly with others. For each item, the percent of other items that item was significantly correlated with was calculated, and this range is shown in Table 27 along with the highest correlation that any one item had with any other item.

Table 27

Inter-Item Correlations for ECERS-R Scores

Scoring Method	Minimum Percent of Items That Correlated with Any Single Item	Maximum Percent of Items That Correlated with Any Single Item	Highest *r*
Traditional	66.70%	100.00%	0.762

In the factor analyses, items were removed from the analysis if more than 10% of the cases had missing data for that item (this only happened for items that were Not Applicable) or if the distribution was highly skewed (more than 2.0 in either direction). As a result, Items 11, 27, and 37 (which may each be Not Applicable) were removed.

In the factor analysis, Bartlett's test of sphericity was significant. Bartlett's test looks at whether the given correlation matrix is an identity matrix, which would mean that items had no relationship with each other. In order for a factor analysis to be appropriate, this test must reject the null hypothesis. Kaiser's criterion held for the analysis; that is, the sample size exceeded 250 and the average communality was greater than .600, so all factors with eigenvalues above 1.0 were initially retained and then compared in the Rotated Component Matrix. Factor loadings less than .400 were suppressed in the Rotated Component Matrix. Components with at least two items loading on them (with loadings greater than .400) were considered to be factors in the solution. Table 28 shows the factor solution, eigenvalues, and associated percent of variance explained. The analysis revealed a 6-factor solution with a total percent of explained variance in scores of over 60%. There were no items that did not load on any of the factors with a loading of at least .400. There were nine items that loaded on two different factors with loadings of greater than .400. In those cases, that item was assigned to the factor on which it loaded the highest.

Table 28

Factor Solution for ECERS-R

Factor	Eigenvalue	Percent of Variance Explained	N of Items Loading/ Total Items in Analysis
Factor 1	7.71	22.67	15
Factor 2	4.51	13.26	8
Factor 3	2.59	7.61	4
Factor 4	2.56	7.54	3
Factor 5	1.86	5.48	2
Factor 6	1.31	3.86	2
Total 6-Factor Solution		60.43	34/34

Regarding which items loaded on which factors and with what loadings, the results are displayed in Table 29. Also included in Table 29 is the developers' subscale placement of the items in the ECERS-R, indicated by items of different colors. Items in green correspond to the Space and Furnishings subscale, items in blue correspond to the Personal Care subscale, items in purple correspond to the Language-Reasoning subscale, items in black correspond to the Activities subscale, items in red correspond to the Interactions subscale, and items in pink correspond to the Program Structure subscale. Though the factor solution included six factors, the items that loaded on those factors did not lay out in exactly the same way that those items were organized into ECERS-R subscales. In other words, despite the similarity between the number of factors and the organization of the ECERS-R into subscales, when the items under those factors and subscales are considered, the similarity is somewhat less apparent. For the ECERS-R subscales of Activities and Interactions, all of the items in the original scale remained together, but the items did not represent the only ECERS-R subscale items in their factors. For the other four subscales of the ECERS-R, items did not group in exactly the same way the developers organized them, though the organization was similar.

Table 29

Factors on which ECERS-R Items Loaded

Item	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Child-related display	0.539					
Music/movement	0.547					
Sand/water	0.552					
Free play	0.555					
Promoting acceptance & diversity	0.558					
Group time	0.574					
Encouraging children to communicate	0.645					
Books & pictures	0.670					
Blocks	0.690					
Dramatic play	0.697					
Furniture for relaxation	0.711					
Fine motor	0.740					
Nature/science	0.746					
Art	0.755					
Math/number	0.761					
Greeting/departing		0.443				
Using language to develop reasoning		0.498				
Supervision of gross motor activities		0.522				
General supervision of children		0.635				
Informal use of language		0.659				
Interactions among children		0.770				
Staff-child interactions		0.773				
Discipline		0.776				
Safety practices			0.592			
Meals/snacks			0.609			
Health practices			0.712			
Toileting/diapering			0.761			
Space for gross motor				0.695		
Gross motor equipment				0.771		
Schedule				0.508		
Room arrangement					0.610	
Space for privacy					0.695	
Indoor Space						0.563
Furniture for care, play, learning						0.752
Nap/rest	(N/A)					
Use of TV, radio, computers	(N/A)					
Provision for children with disabilities	(N/A)					

VU-ECERS temporal stability. The psychometric properties of the smaller number of items that comprise the derived VU-ECERS version of the scale were looked at with the same analyses that were used to analyze the psychometrics of the standard instrument in this chapter. First, the reliability of the new version of the instrument over time was examined. Table 30 displays the correlations between VU-ECERS scores from classrooms that were observed twice in the same academic year when the lead teacher was the same. Correlations are given at the total and subscale levels. All correlations were significant. The Tennessee classes from the PCER study that were observed more than once were not able to be scored on the VU-ECERS due to the fact that scoring involved the deletion of specific indicators, a level for which we have no information from that group of observations. Therefore, the analysis that looked at Tennessee and Missouri classrooms separately regarding the temporal stability of the ECERS-R scores and found different results for each subgroup of classrooms was not able to be conducted with VU-ECERS data.

In comparison to the temporal stability of the original instrument, the total stability correlations for the VU-ECERS are higher for the total and all subscales except for Space and Furnishings. Though differences in the correlations appear small (ranging from .06 to .15), the practical significance of those correlations is that 17% more of the variation in total ECERS-R scores at the second time point can be explained by scores at the first time point for the VU-ECERS than when the same analysis is looked at for the ECERS-R. This suggests that the original instrument contains some indicators and/or items that seem to contribute instability to the total score, and when those indicators/items are removed, the temporal stability is strengthened.

Table 30

Temporal Stability of Total and Subscale VU-ECERS Scores (N = 35)

Source	N of Classrooms	Pearson *r*	p-value
Space and Furnishings	35	0.471	0.004
Personal Care Routines	35	0.732	0.000
Language-Reasoning	35	0.651	0.000
Activities	35	0.679	0.000
Interaction	35	0.482	0.003
Program Structure	35	0.769	0.000
Total Score	35	0.776	0.000

The temporal stability of VU-ECERS data at the item level was also examined. Table 31 shows the correlations between two time points for VU-ECERS item scores in classrooms that were observed twice in the same year when the lead teacher was the same. Nonsignificant correlations are bolded. At the item level, 82% of the significant correlations were significant at the .01 level (two-tailed). A lesser portion, 18%, were significant at the .05 level (two-tailed). The significant correlations ranged from .382 to .719. Of those correlations, 35% of them were less than .500, meaning that over a third of the items included that were significantly correlated from one time point to the next, only 25% or less of the variation in scores at Time 2 could be predicted by scores at Time 1. The item-level correlations over time are often lower than the correlations at the subscale and total levels, suggesting that while some items are scored differently from one time point to the next, the total score is not very affected by those differences.

Table 31

Temporal Stability of Item-Level VU-ECERS Scores (N = 35)

Item	*r*
Indoor Space	0.490
Furniture for care, play, learning	(Removed)
Furniture for relaxation	(Removed)
Room arrangement	(Removed)
Child-related display	(Removed)
Space for privacy	(Removed)
Space for gross motor	(Removed)
Gross motor equipment	**0.259**
Greeting/departing	**0.234**
Meals/snacks	0.501
Nap/rest	(Removed)
Toileting/diapering	0.719
Health practices	**0.254**
Safety practices	0.693
Books & pictures	0.515
Encouraging children to communicate	0.382
Using language to develop reasoning	0.500
Informal use of language	0.516
Fine motor	**0.270**
Art	(Removed)
Music/movement	(Removed)
Blocks	(Removed)
Sand/water	(Removed)
Dramatic play	0.384
Nature/science	0.505
Math/number	0.416
Use of TV, radio, computers	(Removed)
Promoting acceptance & diversity	**0.238**
Supervision of gross motor activities	**0.222**
General supervision of children	0.451
Discipline	0.449
Staff-child interactions	0.517
Interactions among children	**0.315**
Schedule	0.682
Free play	0.686
Group time	0.687
Provision for children with disabilities	**-0.436**

Note. Nonsignificant values are bolded.

Policy implications of observation time for VU-ECERS scores. Regarding temporal stability, this study also examined changes in a classroom's standing relative to the quality funding cut-score when observed at different times of the academic year. Table 32 displays the results of the McNemar two-tailed nonparametric test that was used to examine whether the extent of change from one side of the cut-score to the other across time was significant with VU-ECERS data. The test did not reveal that significant change occurred from one time point to the next.

Table 32

Temporal Stability of Total VU-ECERS Scores Relative to Funding Quality Cut-Scores

Scoring Method	Classrooms Below the Cut-Score At Both Times	Classrooms Above the Cut-Score At Both Times	Classrooms Changing from Below to Above	Classrooms Changing from Above to Below
Traditional	6	20	6	3

Classroom characterization changes from observation time change for VU-ECERS scores. Table 33 displays the results of the Wilcoxon matched pairs signed-ranks two-tailed nonparametric test that was used to examine whether the extent of change in categories across time was significant. The test did not reveal that significant change occurred from one time point to the next.

Table 33

Temporal Stability of Total VU-ECERS Scores Relative to ECERS-R Quality Categories

Scoring Method	Classrooms Not Changing	Classrooms Moving Down	Classrooms Moving Up	Z
Traditional	22	6	7	-0.277

Observation length and start time as predictors of VU-ECERS scores. Next, the relationship between length of observation time and VU-ECERS scores was examined. Across the 263 classrooms that had information on observation length and had enough data to score the VU-ECERS, the mean observation length was 4.19 hours with a standard deviation of 1.36 hours. The minimum observation length was 1.92 hours and the maximum length was 7.17 hours. Table 34 shows the results of the linear regressions run with observation length as the predictor and total and subscale scores as dependent variables. With the exception of Personal Care Routines, all of the significant *t*-values were negative, indicating that longer observation periods predicted lower VU-ECERS scores. The same analyses with ECERS-R scores (shown in Table 24) displayed the same pattern of results regarding the negative and positive *t-values*. As with the regression predicting total and subscale ECERS-R scores from observation length, length significantly predicted variation in all VU-ECERS scores except for the Personal Care Routines and Language-Reasoning subscale scores.

Table 34

Predicting Scores from Observation Length Using VU-ECERS

Source	Standardized β	*t*-value	p-value
Space and Furnishings	-0.339	-5.813	0.000
Personal Care Routines	0.086	1.392	0.165
Language-Reasoning	-0.099	-1.600	0.111
Activities	-0.145	-2.374	0.018
Interaction	-0.160	-2.625	0.009
Program Structure	-0.355	-6.137	0.000
Total Score	-0.230	-3.820	0.000

Along with observation length, the start time of the observation was also examined as a predictor of VU-ECERS scores. The earliest observation began at 7:00 a.m., while the latest observation began at 12:45 a.m. The average start time was 8:39 a.m. (SD=1.17 hours). Like in the same analysis with the ECERS-R data, the start time of the observation was not a significant predictor of VU-ECERS scores.

Internal consistency of VU-ECERS. The internal consistency of the VU-ECERS subscales was examined by calculating Cronbach's Alpha for the entire scale and examining the effects on that Alpha if a subscale was removed. The correlations of subscales to total scores were also examined. Table 35 displays the Alpha, change in Alpha, and item to total correlations. As in the ECERS-R analyses regarding internal consistency, the overall Alpha was above .80. The overall Alpha was decreased with the removal of every subscale, but it never decreased by more than .047. All subscales had moderate to high correlations with the total scale, with correlations ranging from .669 to .839.

Table 35

Internal Consistency for ECERS-R Subscales with VU-ECERS Data

Source	Alpha (based on Standardized items)	Decrease in Alpha if Subscale is Deleted	Subscale to Total Correlation
Total Scale	0.863		
Space and Furnishings		0.002	0.669
Personal Care Routines		0.018	0.720
Language-Reasoning		0.044	0.823
Activities		0.024	0.731
Interaction		0.047	0.839
Program Structure		0.047	0.835

Note. All correlations were significant at the .01 level (two-tailed).

Concurrent validity of VU-ECERS. Factor analyses of the items on the VU-ECERS were conducted using a principal component extraction method with varimax rotation. First, inter-subscale correlations were examined (displayed in Table 36). While all correlations were significant, they were not high enough to indicate that subscales were duplicative.

Table 36

Inter-Subscale Correlations for VU-ECERS Scores

Scoring Method	N	Mean *r*	Correlation Range
Traditional	272	.513	.310-.722

Note. All correlations were significant at the .01 level (two-tailed).

Next, the inter-item correlations were examined to check for items that did not correlate with many of the other items or items that correlated too highly with others. For each item, the percent of other items that item was significantly correlated with was calculated, and this range is shown in Table 37 along with the highest correlation that any one item had with any other item.

Table 37

Inter-Item Correlations for VU-ECERS Scores

Scoring Method	Minimum Percent of Items Significantly Correlated With	Maximum Percent of Items Significantly Correlated With	Highest *r*
Traditional	66.70%	100.00%	0.774

In the factor analyses, items were removed from the analysis if more than 10% of the cases had missing data for that item (this only happened for items that were Not Applicable) or if the distribution was highly skewed (more than 2.0 in either direction). As a result, Item 37 was removed due to missing data.

In the factor analysis, Bartlett's test of sphericity was significant. Kaiser's criterion held; the sample size exceeded 250 and the average communality was greater than .600, so all factors with eigenvalues above 1.0 were initially retained and then compared in the Rotated Component Matrix. Factor loadings less than .400 were suppressed in the Rotated Component Matrix. Components with at least two items loading on them (with loadings greater than .400) were considered to be factors in the solution. Table 38 shows the factor solution, eigenvalues, and associated percent of variance explained. The analysis yielded a five-factor solution with a total explained variance of 62.2%, slightly higher than the 60.43% variance explained in the ECERS-R with the six-factor solution. No items did not load on any of the factors with a loading of .400

or greater. Nine items loaded on more than one factor with a loading of greater than .400. Those items were assigned to the factor on which they loaded the highest.

Table 38

Factor Solution for VU-ECERS

Source	Eigenvalue	Percent of Variance Explained	N of Items Loading/ Total Items in Analysis
Factor 1	4.19	17.47	9
Factor 2	3.23	13.45	5
Factor 3	2.91	12.13	3
Factor 4	2.51	10.44	4
Factor 5	2.09	8.72	3
Total 5-Factor Solution		62.20	24/24

Regarding which items loaded on which factors, the results are displayed in Table 39. Also included in Table 39 is the developers' subscale placement of the items in the original ECERS-R organization, indicated by color. Items in green correspond to the Space and Furnishings subscale, items in blue correspond to the Personal Care subscale, items in purple correspond to the Language-Reasoning subscale, items in black correspond to the Activities subscale, items in red correspond to the Interactions subscale, and items in pink correspond to the Program Structure subscale. For the Personal Care Routines, Language-Reasoning, and Activities subscales, many of the included items did hang together. However, regarding the other three subscales of the ECERS-R, items on the VU-ECERS did not group similarly to how the developers organized them.

Table 39

Factors on which VU-ECERS Items Loaded

Item	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Using language to develop reasoning	0.504				
Group time	0.525				
Fine motor	0.575				
Books & pictures	0.587				
Promoting acceptance & diversity	0.604				
Encouraging children to communicate	0.614				
Nature/science	0.722				
Math/number	0.767				
Greeting/departing		0.445			
Informal use of language		0.587			
Interactions among children		0.704			
Discipline		0.707			
Staff-child interactions		0.801			
Dramatic play			0.641		
Gross motor equipment			0.769		
Schedule			0.773		
Indoor Space				0.438	
Meals/snacks				0.608	
Toileting/diapering				0.816	
Health practices				0.721	
Safety practices				0.567	
Supervision of gross motor activities					0.800
General supervision of children					0.644
Free play					0.518
Furniture for care, play, learning	(Removed)				
Furniture for relaxation	(Removed)				
Room arrangement	(Removed)				
Space for privacy	(Removed)				
Child-related display	(Removed)				
Space for gross motor	(Removed)				
Nap/rest	(Removed)				
Art	(Removed)				
Music/movement	(Removed)				
Blocks	(Removed)				
Sand/water	(Removed)				
Use of TV, radio, computers	(Removed)				
Provision for children with disabilities	(Removed)				

Factor scores for classrooms using the ECERS-R were correlated with factor scores for classrooms using the VU-ECERS. Correlations are shown in Table 40. Factor scores from each of the first five factors from the ECERS-R factor solution correlated highly and positively with factor scores from exactly one of the five factors from the VU-ECERS factor solution, and vice-versa, with no correlation lower than .686. Factor scores from the sixth factor of the ECERS-R solution did not correlate highly with any factor scores from the VU-ECERS. Factor scores from four ECERS-R factors and all VU-ECERS factors correlated significantly with more than one factor with less strength, and two of those significant correlations were negative, but none of those correlations were above .383. Generally, Factors 1, 2, and 5 from the ECERS-R were most highly correlated with their corresponding factors from the VU-ECERS. ECERS-R Factor 3 was most highly related to VU-ECERS Factor 4, and ECERS-R Factor 4 was most highly related to VU-ECERS Factor 3. When considering the types of items that loaded onto each factor, the factor scores for factors that were most highly correlated involved factors with similar types of items across the two instrument versions. The high factor score correlations speak to the concurrent validity of the VU-ECERS and the ECERS-R.

Table 40

Correlations Among Factor Scores for ECERS-R and VU-ECERS

			VU-ECERS Factor		
ECERS-R Factor	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Factor 1	0.865**	-0.044	0.383**	0.019	-0.027
Factor 2	0.122*	0.854**	-0.015	0.003	0.356**
Factor 3	0.072	-0.036	-0.044	0.944**	-0.005
Factor 4	-0.222**	0.084	0.845**	0.016	0.064
Factor 5	-0.069	-0.113	0.042	0.001	0.686**
Factor 6	0.035	0.219**	-0.128*	0.220**	-0.090

**Note.* Correlation is significant at the .05 level (2-tailed).
***Note.* Correlation is significant at the .01 level (2-tailed).

Research Question 3 Summary

This question examined the psychometric properties of the ECERS-R and the VU-ECERS, a version of the ECERS-R that contained indicators that field experts agreed were important to quality. The analyses for the two instrument versions were very similar, with only slight differences in statistics. Both methods were shown to be reliable across time for total scores and four out of the six subscale scores, with VU-ECERS scores having slightly higher correlations than ECERS-R scores. With both instruments, the length of observation was a significant predictor of total scores, with longer observation periods predicting lower quality scores. The internal consistency of both instrument versions was high. In the factor analyses, though the number of factors in the factor solutions of each method was different, over 60% of the variance in scores was explained by the factor solution accepted for each instrument version. High correlations of factor scores between instrument versions showed concurrent validity for the scale versions.

Research Question 4: Using the VU-ECERS, to what extent are the results affected by alternative scoring methods?

To answer this question, the VU-ECERS data were analyzed in the same way as the ECERS-R data in Research Question 1. Data from the VU-ECERS were compared to data from the ECERS-R when both were scored using the traditional method. The number of classrooms included in the analysis was 272 (classrooms with sufficient data to allow scores to be calculated for both instrument versions). Correlations between instrument versions were run on classrooms’ total scale and subscale scores, displayed in Table 41.

Table 41

Correlations Between VU-ECERS and ECERS-R Subscale and Total Scores

Source	N	VU-ECERS Mean	VU-ECERS SD	ECERS Mean	ECERS SD	*r*
Space and Furnishings	274	4.56	1.72	4.49	1.15	0.714
Personal Care Routines	274	3.92	1.61	3.76	1.57	0.981
Language-Reasoning	274	4.71	1.36	4.71	1.37	0.998
Activities	274	4.26	1.23	3.90	1.23	0.894
Interaction	274	5.04	1.82	5.00	1.82	0.998
Program Structure	274	4.86	1.75	4.81	1.75	0.988
Total Score	274	4.56	1.22	4.46	1.21	0.982

Note. All correlations are significant at the 0.01 level (two-tailed).

All correlations between VU-ECERS and ECERS-R at the subscale and total score levels were positive, significant, and high. Correlations ranged from .714 to .998 for subscale scores, with the highest correlation corresponding to the Language-Reasoning and Interaction subscales. The lowest subscale correlation between instrument version scores was for Subscale 1, Space and Furnishings.

Instrument score comparison. Data from the VU-ECERS was compared to the same observational data from the ECERS-R. Classrooms were ranked on their standing relative to each other with each of the instrument versions, and comparisons were made using correlational analysis and examination of rank movement. The correlation between classroom rankings from the two instrument versions for 272 classroom observations was .980 (p <.001). While the correlation was very high, there was substantial movement in relative ranking for classrooms when the instrument version was altered, though the extent of movement for the majority of classrooms was slight. In the comparison, only 14 of the classrooms did not change relative ranking. Of the 258 classrooms that did move, 54% of them only moved 10 slots or less, although 19 classrooms (7.4%) moved 30 slots or more in either direction.

Policy implications of scoring methods. The change in a classroom's location above or below the quality funding cut-score when the instrument version was altered was examined. Table 42 shows the amount of change around that cut-score when classroom data on the ECERS-R was compared to classroom data on the VU-ECERS. The McNemar two-tailed nonparametric test revealed that a significant number of classrooms changed their location relative to the funding cut-score in the instrument comparison.

Table 42

Classrooms Changing Location Relative to the Quality Funding Cut-Score with VU-ECERS Data

Source	Classrooms Below the Cut-Score With Both Versions	Classrooms Above the Cut-Score With Both Versions	Classrooms Changing from Below to Above	Classrooms Changing from Above to Below
ECERS-R v. VU-ECERS	94	162	13	3

Classroom characterization by instrument version. The change in a classroom's categorization relative to the four ECERS-R quality categories of Inadequate, Minimal, Good, and Excellent when the instrument version was altered was examined. Table 43 shows the number of classrooms that moved to another category when the VU-ECERS instrument was used in place of the ECERS-R. The Wilcoxon matched pairs signed-ranks nonparametric test revealed significant category change. For every classroom that changed categories, whether up or down, no classroom moved more than one category in either direction. Most of the classrooms that changed moved from Minimal with the ECERS-R to Good with the VU-ECERS.

Table 43

Classrooms Changing ECERS-R Quality Categories with VU-ECERS Data

Source	Classrooms Not Changing	Classrooms Moving Down	Classrooms Moving Up	Z
ECERS-R v. VU-ECERS	245	4	23	-3.66

Note. Statistic was significant at the .001 level (2-tailed).

Research Question 4 Summary

This question examined the policy implications of using the VU-ECERS as compared to the ECERS-R. High correlations were found between instrument versions for both subscale and total scores. A very high correlation was found between classrooms' relative ranking from total ECERS-R scores and classrooms' relative ranking from total VU-ECERS scores, though some movement in ranking did exist. As with the ECERS-R comparisons of scoring method, there was a significant amount of change in classrooms' standing relative to the quality cut-score when the instrument version was altered, as well as change in classrooms' standing relative to ECERS-R quality categories. The majority of classrooms that changed their relative standing in regards to cut-scores or quality categories moved to a more positive position (higher than the cut-score or to a higher quality category) when the VU-ECERS was used as opposed to the ECERS-R.

The total scores are stable across instrument versions; the correlations between scores using different versions are very high. However, as discussed in Research Question 1, although the majority of classrooms that moved relative ranking when the VU-ECERS was used did not move very far, the implications of that movement are more evident when the potential impact on funding is considered. When examining the effect of the alternate instrument version on a classrooms standing relative to a quality cut-score that is used in Tennessee, almost 5% of classrooms that would’ve been under that cut-score with the ECERS-R would have been eligible (aside from the other variables that are involved in determining funding) for funding when the VU-ECERS was used.

CHAPTER V

SUMMARY, DISCUSSION, AND CONCLUSIONS

The objectives of this study were to critically examine the ECERS-R using four broad questions. First, the study looked at how different scoring methods applied to the same data might influence the final ECERS-R quality ratings that a classroom achieves. Second, the study examined how the ECERS-R reflects the current field view about what aspects of a classroom contribute to quality and how those aspects should be organized. Third, the psychometric properties of the ECERS-R and a new version of the instrument comprised of aspects of a classroom that field experts consider important to quality were examined. And fourth, this study sought to examine how scores from a new version of the ECERS-R might affect policy decisions. This final chapter provides a summary of the analytical results, a discussion about the implications of the findings, and an acknowledgement of the strengths and limitations of the research.

Summary of Results

Research Question 1

This question examined how the results of the ECERS-R were affected by using different scoring methods on the same data. The traditional method of ECERS-R scoring was compared to a summative method that summed all indicators that were positively-scored for an observation, and to a summative method that summed all indicators that were positively-scored up to the spot on each item that the stop-rule was put into place. Originally each summative method was looked at in two different ways, one that included indicators from the ECERS-R that could have been scored as Not Applicable, and one that did not include such indicators. This allowed for ECERS-R scores from five scoring methods to be compared. The relative ranking of classrooms to each other was compared when the scoring method was altered. Scoring methods that altered the presence or absence of Not Applicable indicators showed such similar results to their counterparts that the rest of the analyses were conducted with the summative methods that excluded those indicators. Across all comparisons, correlations between rankings were high. Despite the high correlations, there was a lot of movement up and down in ranking when the scoring method changed, although the majority of that movement was over small portions of the range of rankings.

There was also significant movement around a funding cut-score when Tennessee's example of 4.0 was used as the funding determinant score, as well as movement among ECERS-R quality categories when the scoring method of the ECERS-R was altered. Furthermore, the pattern of movement was not the same across all scoring method comparisons. Changing from either the traditional method or the summative-stop method to the summative-no stop method always showed classrooms moving to a more positive spot (either from below the funding cut-score to above it or from a lower ECERS-R quality category to a higher one). However, changing from either the traditional method or the summative-no stop method to the summative-stop method always showed classrooms moving to a more negative spot.

Research Question 2

This question examined how much agreement existed among field experts and ECERS-R developers concerning the organization and content of the ECERS-R. Sixteen experts in the fields of early childhood and early education completed surveys in which they indicated the subscale under which they felt each ECERS-R indicator should be placed. Experts also indicated the extent of their agreement that each indicator was an important component of their personal definition of quality. For both portions of the survey, experts were allowed to indicate agreement on subscale placement and indicator importance with two options, "A great deal" and "Somewhat." There was a high degree of agreement on ECERS-R organization when the two agreement categories were collapsed. When just the highest degree of agreement was examined, experts agreed with developers on the placement of almost two-thirds of the indicators. Differential agreement was seen at the level of the ECERS-R subscales. The ECERS-R subscales of Personal Care Routines, Activities, and Program Structure had the lowest mean percentages of indicators on which subscale placement was agreed upon by experts and developers. The subscale for which the mean percent of indicators agreed upon by experts and developers was the highest out of the six subscales was Space and Furnishings. There was more disagreement among experts than there was between experts and developers on the organization of the indicators.

In terms of agreement between experts and ECERS-R developers on the content of the instrument, the panel agreed that most indicators were important to quality at some level. However, there was more disagreement among experts about how important they felt the indicators were to their definitions of quality ("A great deal" or "Somewhat"). With the two agreement categories collapsed, over half of the ECERS-R indicators were considered important to quality by all 16 experts. When the greatest degree of agreement was examined separately, fewer than 10% of the indicators were considered important by all experts. At the subscale level, subscales with indicators pertaining more to interactions in the classroom had the highest percentage of indicators that at least half of the panel agreed were of great importance to quality. The percent of indicators agreed upon by at least half of the panel was much smaller for ECERS-R subscales whose indicators pertained to aspects of the physical environment of the classroom. Most of the indicators that all experts did not agree were of great importance to quality were at the higher anchors of the ECERS-R scale, as opposed to those indicators that represent Inadequate or Minimal quality.

Analyses from this research question led to the creation of a new version of the ECERS-R, called the VU-ECERS, that was comprised of those indicators that at least half of the experts agreed were of great importance to quality. Any items with significant numbers of indicators that did not meet this criterion were removed. In total, 13 entire items were deleted from the ECERS-R to create the VU-ECERS.

Research Question 3

This question examined the psychometric properties of the ECERS-R and the VU-ECERS. In general, analyses related to the reliability and validity of the instruments showed that both versions had fairly sound properties. Internal consistency was high for both versions. Sixty classrooms were observed two times in the same year with the same lead teacher at both times. Analyses with those classrooms revealed that scores from all subscales and totals for both instrument versions were reliable over time when classroom scores at one point in the year were correlated with the same classroom scores at another point in the year. However, not all items showed stability over time. Despite the significance of the correlations from one time point to the next, when the percent of variance explained was considered, scores at the first point explained less than half of the variation in total scores at the second time point for the ECERS-R data, and that amount of variation explained was higher than the amount explained by any of the subscales. For the VU-ECERS, the percentages were slightly higher, with 60% of variation in time two total scores explained by time one total scores. There was no significant change of classrooms around the quality cut-score or among ECERS-R quality categories across time with either the ECERS-R or the VU-ECERS data. However, data from one of the secondary data sets did suggest that scores may not be stable across time, and significant change around quality cut-scores or among ECERS-R quality categories from one time point to the next may be significant.

With both instruments, the length of observation was a significant predictor of total scores, with longer observation periods predicting lower quality scores. Shorter observation length was also a significant predictor of higher scores on both instrument versions for all subscales except for Personal Care Routines and Language-Reasoning, which were not significantly predicted by observation length for either of the two instruments.

Factor analyses of the two instrument versions revealed similar factor structures. Items from the Activities subscale tended to hang together in the solutions for both instruments. Despite a six-factor solution for the ECERS-R and a five-factor solution for the VU-ECERS, the analyses did not lend empirical support for the six subscale designations that the ECERS-R developers created. The VU-ECERS demonstrated concurrent validity with the ECERS-R with high and significant correlations among factor scores from the two instrument versions.

Research Question 4

This question examined the policy implications of using the VU-ECERS as compared to the ECERS-R. Scores for both instrument versions at the total and subscale levels were highly correlated. The same pattern of results found in Research Question 1 for scoring method comparisons was found for instrument version comparisons in this question's analyses. A high correlation was found between classrooms' relative ranking from total ECERS-R scores and classrooms' relative ranking from total VU-ECERS scores, though some movement in ranking did exist. Significant change around the quality cut-score and among ECERS-R categories was found when the instrument version was altered. When ECERS-R scores were replaced with VU-ECERS scores, most of the classrooms that moved around were moved to a more positive spot (above the cut-score or to a higher quality category). Regarding specific policy implications, almost 5% of the classrooms included in the analysis that would not have been funded based on their ECERS-R scores would have been eligible to receive funding with their VU-ECERS scores.

Issues

The ECERS-R Scoring and Observation Requirements

As a result of the analyses in this paper, several issues arose concerning the ECERS-R itself, specifically related to scoring. In the instrument's instructions, ECERS-R developers briefly mention the alternative scoring procedure of using each individual indicator. If such alternatives are not only possible but included in the tool's instructions, the consequences of using such methods must be researched and provided to the users. The developers do mention that scoring a classroom using all of the indicators in the ECERS-R requires a longer observation period. They do not mention, however, the impact that alternative scoring methods may have on the resulting ECERS-R scores. One example where the impact of alternative scoring methods could be of concern relates to how the total score is calculated. Because the developers included different numbers of indicators under each item, some items will carry more weight in the final score if a summative method is used without consideration for the numbers of indicators under each item. ECERS-R developers also fail to mention the expectations of the instrument performance under alternative scoring methods. This research offers information about the psychometrics of the ECERS-R across three scoring methods.

This study demonstrates that characteristics of the observation itself can influence the resulting ECERS-R scores, particularly regarding the length of the observation period. In this research, the length of time an observer was in the classroom was a significant predictor of ECERS-R scores. Shorter observation periods resulted in higher scores, despite the fact that the mean observation length for the classrooms included in this study was longer than the length recommended by the ECERS-R developers. Measures that purport to assess quality through direct observation need to have more stringent requirements of the users, particularly concerning aspects that can affect the outcome.

The negative relationship obtained between observation length and scores might be explained by two factors. First, during an observation in a classroom it is hard to see many of the characteristics examined by ECERS-R indicators. For such types of indicators, developers recommend that teacher interviews be conducted to collect enough information to score them. Observation periods that are shorter would very likely increase the number of indicators that would need to be scored through a teacher interview as opposed to direct observation. A heavier reliance on teacher report in shorter observation periods is likely to positively bias the scores. It might be helpful to users if notes were added to specify those indicators that were scored solely on the basis of teacher report. A second explanation is the fact that shorter observation periods might lead to the observation of fewer behaviors that would cause indicators to be negatively scored. For instance, an observer who assesses one free play activity with many conversations between the teacher and children might conclude that the indicator corresponding to such behavior could be positively scored. However, if the observer was in the classroom for a longer period of time and witnessed several additional periods of free play activity in which no conversations were had, the corresponding indicator might not be positively scored. Regardless of the explanation, the influence that observation length has on ECERS-R scores should not be ignored by developers or users.

The individual indicators, although they arguably give a more detailed picture of a classroom when examined than do the items or subscales, have led to issues with the ECERS-R. When developers revised the original version of the instrument to create the ECERS-R, what had previously been general descriptions of classrooms under each item anchor became individual indicators that were to be checked if present, and the number of indicators scored positively was linked to an item score on a one to seven scale. Although creating these behavioral indicators allowed for higher inter-observer reliability, the new indicator-based scale version had new issues that were not applicable to the original version of the instrument. The ECERS-R, based on indicators, could be said to function well neither left to right, nor top to bottom. If the ECERS-R functioned left to right within an item, each item would include indicators that were represented under each scale anchor with increasing degrees of quality. Indeed, for some items there are ECERS-R indicators that do appear at every level of the item. For example, in the Greeting/Departing item, indicators pertaining to how and why parents bring their children to the classroom are gradated throughout the entire scale, represented at each of the four quality levels. However, this is not the case with every item. For some items their indicators pertain to behaviors that are only represented at either the low or high end of the item. For example, in the Fine motor item, an indicator concerning whether materials for different levels of difficulty are accessible to the children is only found under the “Good” anchor but does not appear in any other form under any of the other anchors on the scale.

In addition, if the ECERS-R functioned top to bottom within an item, an observer would score the indicators under the anchor, beginning with the topmost indicator and moving towards the bottom most indicator, until a negative score was awarded to an indicator. At that point, one could assume that a classroom represented that level of quality on the item and no higher. It would not be possible for a classroom that scored a “3” on an item to have indicators under higher anchors that could have been positively-scored. As previously discussed in this paper and shown in the analyses, this assumption does not fit the ECERS-R.

Policy Concerns

This study showed that ECERS-R scores seem to be relatively unaffected by the use of alternative scoring methods as far as the rank order of the classrooms. The instrument displayed fairly sound psychometric properties, demonstrating score stability across time within a year, relatively high internal consistency, and relatively high inter-rater reliability. However, issues concerning policy arise when ECERS-R scores are considered in relationship to specific benchmarks and funding is distributed according to that consideration. The issue for policy decisions lies in the movement of classrooms across benchmarks when the scoring method or time point of observation is altered. Policy makers must be aware that a program receiving funding based on its ECERS-R scores might not have been eligible for those funds if the ECERS-R had been scored in a different way, if, as noted above, the length of the observation time was increased, or if the observation had been made at a different time during the year. These issues should be of particular interest not just for policy makers but for employees at programs in states where ECERS-R scores are already tied to funding decisions. It is the livelihood of those employees, after all, that is directly affected by quality scores in states such as Tennessee. If higher ECERS-R scores mean higher reimbursement rates, and ECERS-R scores can be influenced not only by classroom characteristics but by characteristics pertaining to the way the observation itself is conducted, serious consideration should be given to the way in which the ECERS-R is used in the realm of policy.

Defining Quality

In the literature on quality reviewed in this paper, there was specific mention of the difficulty associated with defining what quality is in early childcare and specifically which components of an early educational environment compose that quality. As discussed earlier, quality assessment tools often fail to include the definition of quality held by the developers, putting the onus on the user to infer the developer's definition from the items included in the instrument. The ECERS-R was designed to measure the quality of an environment at a global level. It includes items pertaining to the process quality of a classroom, with process quality represented by items that linked the physical environment of the classroom to the interactions and activities that take place within the environment. Field experts surveyed in this study indicated that certain indicators from the ECERS-R were not as important to their own personal definitions of quality. Most of the indicators that were eventually removed from the ECERS-R to create the VU-ECERS involved the physical environment of the classroom; many of the indicators assessed how the room was arranged, the organization of materials in the classroom, and the number and type of materials available to children. None of the ECERS-R indicators that involved the language environment of the classroom, the type of instruction, or the general interactions between people in the classroom were removed. This suggests that, although the ECERS-R looks at a wide range of indicators to assess the global quality of the classroom (with a heavy emphasis on the physical environment), field experts define quality more in terms of the instruction and interactions that take place within the space.

The type of indicators that were not considered as important to quality by field experts also suggests that the distinction between structural and process quality discussed in the literature review may be an oversimplified dichotomy. Features of structural quality, mainly distal features including teacher wage, education, and training, are clearly separate classroom characteristics that are more influenced by outside forces. Process quality, as it has been discussed in previous research, is not so cohesive in its features. Responses from field experts in this study seem to separate process features that are related to the physical environment of the classroom from process features that involve human interactions. There appears to be the need to add a third distinction to the types of quality, one that is specific to the learning environment and that involves interactions among people, whether it be between teachers and children or among children.

Another issue with the structural and process quality distinction arises when the removed indicators for the VU-ECERS are considered. Field experts in this study were much more likely to disagree that indicators from the highest ECERS-R scale ratings were important to quality than they were to disagree that indicators from the lower end of the ECERS-R were important.

Significantly more indicators under the ECERS-R anchors of Good and Excellent were removed than were indicators from the Inadequate and Minimal categories. Thus, there might indeed be two types of quality but not along the division between structure and process. Instead one type of quality, comprised of indicators at the lower two ECERS-R quality levels, might be viewed as a *floor* of good practice. Indicators under these anchors indicate a classroom that is generally safe and healthy for children, one in which manipulatives are available to children so that learning can occur in the absence of harsh discipline. This type of quality can be thought of as benchmark quality, or the basic quality necessary for any educational environment serving young children. Beyond that floor, another type of quality was comprised of indicators at the higher two ECERS-R quality levels. This type of quality indicates characteristics of a classroom that move it beyond the minimal benchmark quality standards and into the realm of high quality *as individually defined by different groups of developers*. Indicators representing this type of quality pertain to the types of questions teachers ask children, the methods of instruction that teachers use, the expansion of children's ideas, the complexity and variety of academic activities and experiences provided for the children, etc. While everyone appears to be able to agree on an appropriate floor for good practice, there may be considerable disagreement on what makes a classroom of high quality.

In their surveys, field experts often responded with comments about how they felt the ECERS-R indicators represented quality (many of the comments are listed in Appendix C). Five of the sixteen experts specifically talked about feeling as if the ECERS-R was attempting to measure two distinct types of quality. The distinction between benchmark quality and high quality was made repeatedly. One expert responded:

> There are some things I believe are absolutely necessary in any program and I think they have to be in place before you can even start talking about quality. These include safety, health, and kindness; safety and health are usually covered by licensing, and kindness sometimes seems all too rare. So as a parent, these would be essential basics. Then I would start looking for quality . . . But in the rating scale, I couldn't bring myself to say that "dangerous playground/furniture" was not at all related to quality. So my 'very important' category is a combination of apples (safety, health, kindness) and oranges (rich language, challenging activities, content).

Although the ECERS-R uses its categorizations of different points along the scale to distinguish indicators representing minimal quality from indicators representing excellent quality, the field expert survey data suggest that one could have quite different definitions of what constitutes high quality.

The distinction between benchmark and high quality types seems to be influenced by discipline. Most of the field experts surveyed in this study could agree that indicators pertaining to the benchmark quality (at lower levels of ECERS-R quality categories) were important, but opinions began to diverge at higher levels. Experts tended to disagree on what types of characteristics represented the highest quality, and this disagreement was at least in part linked to the experts' areas of specialization. For example, some Good and Excellent indicators pertaining to math and science were rated as less important by experts whose area of specialty was literacy. In other examples, experts with similar specialty areas relating to language and literacy selected different high-level indicators from items involving language as the most important to quality. One practical consequence of this link between expert discipline and quality indicators might be that quality assessment tools include an assessment of benchmark quality but then include an additional tool (or set of tools) tailored to the focus and goals of individual programs, whether the focus is on literacy, language, math, science, social skills, etc.

This idea of a tool kit approach to assessing quality may help with the issue identified in Chapter II that quality has not been consistently linked to child outcomes in previous research. The distinction being suggested here that the field needs to make between different types of quality, either in regard to types of process quality or separating benchmark and high quality, provides a possible explanation for the lack of consistent findings. The problem might lie in researchers' attempts to measure quality globally and generally rather than specifically measuring distinct types of quality more closely linked to the goals of the programs. As discussed previously, studies attempting to link quality to children's achievement have often used the global type of quality measures or combinations of global quality instruments. Specifically, differentiating high quality indicators by discipline or program focus may more conclusively link quality to child outcomes in the areas intended to be affected by the programs.

The suggestion that measures of early childhood environments include indicators that measure what a floor of good practice should involve and also include separate sets of indicators that examine components of high quality in different domains lends itself to a conception of quality assessment tools as pieces of a toolbox. Measures of quality can be thought of as separate tools with unique purposes that can all be included in a kit examining global quality or, if not in a single "kit," then available from curriculum and program developers. If program and curriculum developers were assured that an accepted measure of the floor of good practice existed, they could then concentrate on the critical elements of the program or curriculum that indicated it was being implemented well.

Thus the types of tools needed to assess a general floor of quality might include a measure that looked specifically at health and safety standards often used by licensing agencies, a measure that looked at the materials necessary for learning to occur, a measure that looked at structural indicators of quality like teacher training and education, and so on. In comparison, the types of tools needed to assess quality at higher levels might include a measure that looked at high quality language environments, a measure that looked at rich socio-emotional experiences in a classroom, a measure that looked at high quality in a classroom regardless of content area, etc. Additionally, other measures could be developed to focus on the type of quality that is important to a program's goal. For example, in a Montessori school, a quality assessment might focus on child independence and initiative. In contrast, a quality assessment measure in a Waldorf preschool might focus more on the children's creativity and imagination.

Multiple assessment tools would be useful to users, whether they be program directors, researchers, or policy makers. The user could select the right tool for the purpose of the assessment. Depending on the goal of the user, one or more of the quality assessments in the toolkit they will use might best suit their needs. For example, if policy makers only want to fund programs that reach a certain minimal level of quality, an instrument that assesses the good floor of practice, or benchmark quality that was mentioned earlier, might be all that they would need. If, however, the goal of policy makers is to fund programs that go above and beyond the minimal standards to provide a high quality environment for the children they serve, it is incumbent upon them to define specifically what high quality means. Policy makers could decide that high quality means a general balance among areas related to activities, interactions, materials and program structure, in which case the upper end indicators of the ECERS_R might do.

As discussed above, it is important that developers of assessments purported to measure quality make the definition of quality that the instrument is based on evident to the user. If, as suggested by the expert panel surveyed in this study, quality can be thought of in more than just two dimensions and those dimensions can have specific purposes and focuses, policy makers need to be aware of the specific definition of quality behind an instrument in order to align their purpose to the measure that they choose to use.

The ECERS-R Content

A final issue for the ECERS-R concerns the content. In their comments about the instrument, field experts surveyed in this study mentioned several areas in which they felt the ECERS-R was lacking. Several experts talked about the need to include more items related to teacher planning and assessment. Also, experts suggested that a separate module be developed with indicators pertaining specifically to children with special needs, as the floor of good practice for such children may be higher than one serving the needs of typically-developing children. Several experts mentioned that the instrument did not include any indicators that examined how teachers build on children’s existing knowledge in specific content areas to advance their learning. The issue brought up by almost half of the experts, however, was the insufficient number of indicators pertaining to teacher-child talk and interactions in the classroom. This issue links back to the concern about the wide assortment of characteristics included under the umbrella of process quality.

In sum, The ECERS-R seems to do a good job at assessing components of a classroom that involve the general health and safety of the children. Indicators that focus on those aspects of a classroom are plentiful in the ECERS-R. None of the experts surveyed indicated that more items looking at these characteristics should be included in the instrument. In addition, most of the experts agreed that indicators at the lower levels of ECERS-R quality categories were important to measure. Many of those lower-level indicators examine aspects of a classroom that should be in place for any environment serving young children, but they do not necessarily relate to aspects that indicate high quality. However, the ECERS-R does not do a good job at examining classroom components that look at what most of the experts considered of great importance: teacher-child interactions, instructional content, scaffolding and assessment, rich learning experiences, etc. Surveyed experts did not agree amongst themselves about indicators of quality at the higher ends of the ECERS-R scale. The area of the ECERS-R that is lacking is in its content, which is not an issue solved by altering the scoring method.

Strengths and Limitations

This study's main strength lay in its numbers. With over 250 classrooms in most of the analyses, results are at least comparable as far as sample size is concerned to a few other large studies examining the properties of the ECERS-R. Large sample sizes lower the risk of Type II error by increasing the power to detect effects when they do exist. However, large sample sizes can also cause small effects to appear significant, so it is important that the practical implications of significant results be examined along with the statistics. This study included data from early childhood programs in three different states across different regions of the country. Although state and subgroup differences in quality were not the focus of this study, quality scores were not identical across data sets. Significant results were found that had real implications for policy despite the different levels of quality included in analyses. Another major strength of this study was the inclusion of hundreds of classrooms with ECERS-R data in which every indicator was scored during observation. The majority of studies examining the ECERS-R only include the broader items, subscales, and total levels and most have been limited to scores that were obtained while incorporating the stop-rule. Indicator level data allowed for analyses at the level with the highest inter-rater reliability.

Although the purpose of this study did not involve examining the link between quality and child outcomes, an examination of the prediction of child achievement data from ECERS-R scores calculated with different scoring methods and with alternative versions of the instrument would have contributed to the research concerning links between quality and child outcomes. Many of the analyses included in this study looked at the comparison of scoring methods/ versions of the ECERS-R but did not offer a recommendation as to which method/version was better. It is possible that links between scores and child outcomes could have been different across methods and versions, although a lack of connection to outcome might result from the global nature of the instrument no matter how it was scored.

Another limitation of this study was using secondary data sets as the main source of ECERS-R scores. The use of data sets not originally collected for the direct purposes of this study made certain information unavailable that would have been useful for analyses. For example, an in-depth analysis of inter-rater reliability, specifically looking at how reliability changes across different types of indicators, items, or subscales, was not possible given the information that was provided by collectors of the secondary data sets.

One thing examined in this study that has particular relevance for policy concerns was the temporal stability of the ECERS-R. Because there can be enough of a change in a classroom's score to put it in a different location relative to a possible funding cut-score at one time of the year than it was at another time that year, it means that a program's funding can depend on the time of year that it is observed. Even though the analysis of the temporal stability of ECERS-R scores in the same year yielded significant correlations between scores, there was still a good deal of variation in scores at time two not explained by scores at time one. A limitation of this study was the lack of an expectation of how stable these scores should be expected to be. One could imagine that a classroom's scores on quality measures might increase as the year progresses, owing to stronger teacher-child relationships, child knowledge of and comfort with the daily schedule, child familiarity with behavior management systems, etc. An analysis of the temporal stability of an instrument over the course of a year should be able to account for the amount of expected change in a classroom's scores. However, the difference in scores does not change the policy implications related to the time of year that a classroom is observed and scored for funding consideration.

Additional limitations of this study were related to the Indicator Survey that was completed by field experts. While the small sample size of sixteen experts allowed for a better chance of finding consensus among the panel in regards to the organization and content of the ECERS-R, a larger sample would have perhaps been more representative of the entire field. In addition, the three options provided for experts to indicate their agreement on aspects of the instrument limited the range of responses to a lot of agreement, some agreement, or no agreement at all. More response options would have allowed for experts to better weight aspects that they considered of most importance to quality. Furthermore, the survey was very long and time-consuming to complete because of the need to include all ECERS-R indicators in an evaluation of scale validity.

Conclusions

This dissertation analyzed the properties of the currently most widely-used measure of quality in early childcare environments, the ECERS-R. Analyses examined the implications of the use of alternative scoring methods on ECERS-R data, the extent of agreement among experts in the field as to the organization and the content of the measure, the psychometric properties of the measure in its original form and of an adapted version based on aspects agreed on by the field, and the policy implications associated with using the adapted version. The ECERS-R demonstrated sound reliability and validity in psychometric analyses, specifically related to the internal consistency and temporal stability of the measure. However, the validity of the measure was challenged, with field experts indicating that certain components of the ECERS-R were not important to an assessment of quality. An alternate version of the ECERS-R based around aspects that experts agreed were important quality components showed the same pattern of reliability in psychometric analyses.

Though alternative scoring methods yielded scores that were highly correlated with scores from the traditional scoring method, there was substantial movement in classroom's ranking relative to other classrooms when the scoring method was altered. That movement is of particular concern when considered in light of policy implications. Changing the scoring method with data from the same observation resulted in some classrooms changing their position relative to a cut-score used to determine state funding for programs. Alternative scoring methods also resulted in classrooms changing the ECERS-R quality designation assigned based on those scores. When the version of the ECERS-R was altered to encompass only those items that the field agreed were important to quality, the same policy concerns were seen. Some classrooms changed their location relative to the funding cut-score and ECERS-R quality categories when the alternative version of the ECERS-R was used.

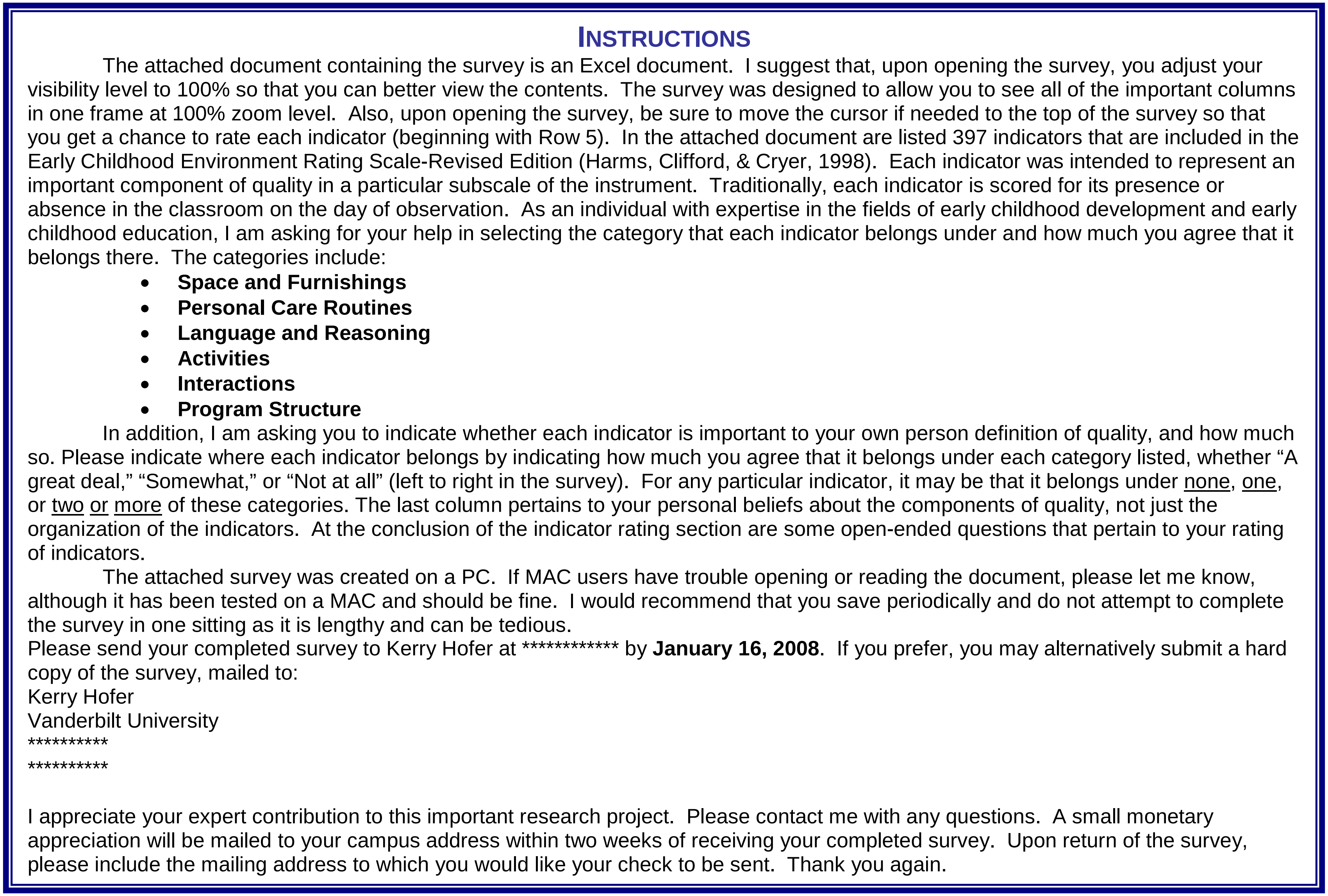
When the ECERS-R is used in policy decisions regarding the funding of programs, certain issues must be considered. A classroom's ECERS-R score is affected not only by the scoring method used and the instrument version used but also by characteristics of the observation. The length of observation and the time of year that an observation occurs can have consequences for the ECERS-R scores that a classroom receives, which can, in turn, have consequences for the funding that a program is eligible to be given. Policy makers who are currently using or are considering using the ECERS-R as a means to assess the quality of programs in their area should attend to the issues raised in this paper.

The main contribution of this research lies in its examination of the content of the ECERS-R. Despite the fact that the psychometric evaluation of the instrument showed it to be a fairly sound measure, there appears to be difficulties with its evaluation of the highest end of the quality dimension in a classroom. This paper began, in part, with an examination of why research indicates that quality is important. No one would disagree that the physical environment of an early childcare environment is important. It is important that children not be put at risk and their health and safety endangered. It can be argued that it is also important that children have some access to materials with which learning can occur. The ECERS-R takes a comprehensive approach to these aspects of a classroom. However, field experts indicated that high quality is not well represented in the ECERS-R. Classroom interactions and learning experiences independent of content are important for children's academic preparation but are not covered in this quality assessment tool. Even if concerns about the scoring method and observation requirements were resolved, the content of the ECERS-R would still be an issue.

The characteristics of a classroom serving young children that are important for those children's academic success may not be assessed in a way that the field can accept. This study suggests that quality measures consist of two general but separate areas, one that looks at what should be in place in a classroom for a good floor of practice, and one that looks at high quality characteristics more narrowly defined by content area.

APPENDIX A

Indicator Survey Instructions and Sample Page



INSTRUCTIONS

The attached document containing the survey is an Excel document. I suggest that, upon opening the survey, you adjust your visibility level to 100% so that you can better view the contents. The survey was designed to allow you to see all of the important columns in one frame at 100% zoom level. Also, upon opening the survey, be sure to move the cursor if needed to the top of the survey so that you get a chance to rate each indicator (beginning with Row 5). In the attached document are listed 397 indicators that are included in the Early Childhood Environment Rating Scale-Revised Edition (Harms, Clifford, & Cryer, 1998). Each indicator was intended to represent an important component of quality in a particular subscale of the instrument. Traditionally, each indicator is scored for its presence or absence in the classroom on the day of observation. As an individual with expertise in the fields of early childhood development and early childhood education, I am asking for your help in selecting the category that each indicator belongs under and how much you agree that it belongs there. The categories include:

- **Space and Furnishings**
- **Personal Care Routines**
- **Language and Reasoning**
- **Activities**
- **Interactions**
- **Program Structure**

In addition, I am asking you to indicate whether each indicator is important to your own person definition of quality, and how much so. Please indicate where each indicator belongs by indicating how much you agree that it belongs under each category listed, whether "A great deal," "Somewhat," or "Not at all" (left to right in the survey). For any particular indicator, it may be that it belongs under none, one, or two or more of these categories. The last column pertains to your personal beliefs about the components of quality, not just the organization of the indicators. At the conclusion of the indicator rating section are some open-ended questions that pertain to your rating of indicators.

The attached survey was created on a PC. If MAC users have trouble opening or reading the document, please let me know, although it has been tested on a MAC and should be fine. I would recommend that you save periodically and do not attempt to complete the survey in one sitting as it is lengthy and can be tedious.

Please send your completed survey to Kerry Hofer at ************ by **January 16, 2008**. If you prefer, you may alternatively submit a hard copy of the survey, mailed to:

Kerry Hofer
Vanderbilt University
****
****

I appreciate your expert contribution to this important research project. Please contact me with any questions. A small monetary appreciation will be mailed to your campus address within two weeks of receiving your completed survey. Upon return of the survey, please include the mailing address to which you would like your check to be sent. Thank you again.

	Indicate the Category that you feel the Indicator belongs under and how much you agree that it belongs there.																		How important is this indicator to *your* definition of quality?		
	Space &			Personal Care			Language-			Activities			Interaction			Program					
Indicator	A great deal	Somewhat	Not at all	A great deal	Somewhat	Not at all	A great deal	Somewhat	Not at all	A great deal	Somewhat	Not at all	A great deal	Somewhat	Not at all	A great deal	Somewhat	Not at all	A great deal	Somewhat	Not at all
	Select one only			*Select one only*			*Select one only*			*Select one only*			*Select one only*			*Select one only*			*Select one only*		
Staff sometimes talk about logical relationships or concepts.	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
Gross motor space has convenient features (close to toilets and drinking water, etc.).	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
No accomodations made for children's food allergies.	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
Space set aside for one or two children to play, protected from intrusion by others.	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
Staff follow through with activities and interactions recommended by other professionals to help children with disabilities meet identified goals.	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
Three-dimensional child-created work displayed as well as flat work.	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
Staff have information for children with disabilities from available assessments.	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
Sufficient furniture for routine care, play, and learning.	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
Children generally follow safety rules.	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
Supervision provided during free play to protect children's health and safety.	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
Staff use greeting and departure as information sharing time with parents.	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□

APPENDIX B

ECERS-R Indicators Removed for the VU-ECERS

Item	**Level**	**Indicator**
Item 1		INDOOR SPACE
	5	Indoor space has good ventilation, some natural lighting through windows or skylights.
	7	Natural light in indoor space can be controlled (blinds or curtains).
Item 2		FURNITURE FOR ROUTINE CARE, PLAY, AND LEARNING
	7	Routine care furniture is convenient to use (cots stored for easy access, etc.).
	7	Woodwork bench, sand/water table, or easel used.
Item 3		FURNITURE FOR RELAXATION
	1	No soft furnishings accessible to children.
	1	No soft toys accessible to children.
	3	Some soft furnishings accessible to children.
	3	Some soft toys accessible to children.
	5	Cozy area accessible to children for a substantial portion of the day.
	5	Cozy area is not used for active physical play.
	5	Most soft furnishings are clean and in good repair.
	7	Soft furnishings in addition to cozy area accessible to children.
	7	Many clean, soft toys accessible to children.
Item 4		ROOM ARRANGEMENT FOR PLAY
	5	At least three interest centers defined and conveniently equipped.
	5	Quiet and active centers placed not to interfere with one another.
	5	Space is arranged so most activities are not interrupted.
	7	At least five different interest centers provide a variety of learning experiences.
Item 5		SPACE FOR PRIVACY
	1	Children not allowed to play alone or with a friend, protected from intrusion by other children.
	3	Children are allowed to find or create space for privacy (behind furniture or room dividers, etc.).
	5	Space set aside for one or two children to play, protected from intrusion by others.
	5	Space for privacy accessible for use for a substantial portion of the day.
	7	More than one space available for privacy.
	7	Staff set up activities for one or two children to use in private space, away from group activities.
Item 6		CHILD-RELATED DISPLAY
	1	Inappropriate display materials for predominant age group (pictures showing violence, materials designed for older children in preschool class, etc.).
	3	Some children's work displayed.
	5	Most of the display is work done by children.
	5	Many items displayed on child's eye level.
	7	Individualized children's work predominates.
	7	Three-dimensional child-created work displayed as well as flat work.
Item 7		SPACE FOR GROSS MOTOR
	5	Adequate space outdoors and some space indoors for gross motor play.
	5	Gross motor space is easily accessible for children in group.
	5	Gross motor space is organized so different types of activities do not interfere with one another.
	7	Outdoor gross motor space has variety of surfaces permitting different types of play.
	7	Gross motor space has convenient features (close to toilets and drinking water, etc.).
Item 8		GROSS MOTOR EQUIPMENT
	1	Very little gross motor equipment used for play.
	5	There is enough gross motor equipment so that children have access without a long wait.
	7	Both stationary and portable gross motor equipment are used.
Item 9		GREETING/DEPARTING
	1	Departure not well organized.
	3	Departure well organized.
	5	Pleasant departure.

Item 10 MEALS/SNACKS

5 Children encouraged to eat independently (child-sized eating utensils provided, etc.).

7 Child-sized serving utensils used by children to make self-help easier.

Item 11 NAP/REST

5 Children helped to relax during nap/rest (cuddly toy, soft music, back rubbed).

5 Nap/rest space is conducive to resting (dim light, quiet).

5 All cots or mats are at least 3 feet a part of separated by a solid barrier.

7 Nap/rest schedule is flexible to meet individual needs (tired child given place to rest in play time).

Item 12 TOILETING/DIAPERING

7 Child-sized toilets and low sinks provided.

Item 13 HEALTH PRACTICES

5 Care given to children's appearance (faces washed, aprons used for messy play).

Item 19 FINE MOTOR

3 Some developmentally appropriate fine motor materials of each type accessible.

5 Many developmentally appropriate fine motor materials of each type accessible for a substantial portion of the day.

5 Fine motor materials are well organized.

7 Containers/accessible storage shelves for fine motor materials have labels and encourage self-help.

Item 20 ART

7 Three-dimensional art materials included at least monthly.

7 Some art activities are related to other classroom experiences.

7 Provisions made for children four and older to extend art activity over several days.

Item 21 MUSIC/MOVEMENET

3 Some music materials accessible for children's use.

3 Staff initiate at least one music activity daily.

3 Some movement/dance activity done at least weekly.

5 Many music materials accessible for children's use.

5 Various types of music are used with the children.

7 Music available as both a free choice and group activity daily.

7 Music activities that extend children's understanding of music are offered occasionally.

7 Creativity is encouraged with music activities.

Item 22 BLOCKS

3 Some clear floor space used for block play.

5 Enough space, blocks, and accessories are accessible for three or more children to build at the same time.

5 Blocks and accessories are organized according to type.

5 Special block area set aside out of traffic, with storage and suitable building surface.

5 Block area accessible for play for a substantial portion of the day.

7 At least two types of blocks and a variety of accessories accessible daily.

7 Blocks and accessories are stored on open, labeled shelves.

7 Some block play available outdoors.

Item 23 SAND/WATER

1 No provisions for sand or water play, outdoors or indoors.

1 No toys to use for sand or water play.

3 Some provisions for sand or water play accessible either outdoors or indoors.

3 Some sand/water toys accessible.

5 Provision for sand and water play (either outdoors or indoors).

5 Variety of sand/water toys accessible for play.

5 Sand or water play available to children for at least 1 hour daily.

Item 24 DRAMATIC PLAY

3 Separate storage for dramatic play materials.

5 Many dramatic play materials accessible, including dress-up clothes.

5 Props for at least two different dramatic play themes accessible daily.

7 Dramatic play props provided to represent diversity (props representing different cultures, equipment used by people with disabilities, etc.).

7 Props provided for active dramatic play outdoors.

7 Pictures, stories, and trips used to enrich dramatic play.

Item 25 NATURE/SCIENCE

3 Children encouraged to bring in natural things to share with others or add to collections (bring fall leaves in from playground, bring in pet).

5 Science/nature materials accessible for a substantial portion of the day.

5 Nature/science materials are well organized and in good condition.

Item 27 USE OF TV, VIDEO, AND/OR COMPUTERS

3 Alternative activities accessible while TV/computer is being used.

5 Computer used as one of many free choice activities.

5 Most of the TV/computer materials encourage active involvement (children can dance, sing, exercise to video; computer software encourages children to think and make decisions).

5 Staff are actively involved in use of TV/video/computer (watch and discuss video with children, help child learn to use computer program).

7 Some of the computer software encourages creativity.

7 TV/computer materials used to support and extend classroom themes and activities.

Item 28 PROMOTING ACCEPTANCE OF DIVERSITY

5 Some props representing various cultures included for use in dramatic play.

Item 29 SUPERVISION OF GROSS MOTOR ACTIVITIES

7 Staff help with resources to enhance gross-motor play (help set up obstacle course for tricycles, etc.).

Item 30 GENERAL SUPERVISION OF CHILDREN

3 Attention given to cleanliness and to prevent inappropriate use of materials (messy science table cleaned up, child stopped from emptying whole glue bottle).

Item 34 SCHEDULE

3 Written schedule is posted in room and relates generally to what occurs.

SUMMARY OF REMOVED INDICATORS:
8 Inadequate (Level 1) indicators removed
17 Minimal (Level 2) indicators removed
40 Good (Level 3) indicators removed
32 Excellent (Level 4) indicators removed

APPENDIX C

Field Expert Comments Concerning the Content of the ECERS-R

Although math is mentioned several times, there is no mention of content knowledge or of the pre/literacy knowledge that is emphasized by funding agencies and state/federal standards.

As I've spent more time in developing countries, I have come to be even less focused on the physical features, schedule, etc. and much more on the features of the interactions between adults and children and the promotion of positive development whatever the methods and furnishings.

I find that when there is something that the ECERS may see as positive but is way at the extreme positive end of the scale I find myself saying 'not at all' [important to quality] because I think the expectation is just unrealistic and not really important. That is in contrast to my reaction to extreme negative, bad stuff, which I am more inclined to say 'very important' [to quality].

I think most of the important indicators are here, but perhaps with too much emphasis on physical environment and not enough on some aspects of cognitive, language, literacy development etc.

I think safety items and anything that might be a licensing concern (handicapped accessibility, nutritional value of foods, basic sanitation that threatens children's safety) should be kept completely separate and reported as a separate index. That way all of the other things would emerge more clearly.

I think that it could be more fine tuned toward actual instructional and activity practices (i.e., describe more what teachers are doing). In addition, I would like more about teacher planning and assessment.

I think that teacher talk/interactions is seriously underrepresented in the scale. The scale represents only very basic aspects of quality, not necessarily high quality.

I think that the scale is measuring quality, albeit at a very minimal level.

Indicators of opportunities for children's sequential early learning--with newer concepts/experiences building upon earlier ones--are not well represented among these indicators.

More than any other indicators, I found myself categorizing [indicators specific to children with disabilities] in many of the ECERS domains. I also found myself clicking 'very important' [to quality] for many, many of these--I think because they almost all seemed essential if kids with disabilities are to have access to the general curriculum and have good outcomes. This is in contrast to many of the ECERS items which seem kind of nice to have or to do, but which one would be hard pressed to say are essential.

Need for more specific items related to the connections between assessment and program planning; more detailed items related to responsiveness of adults to children's needs and links between curriculum and assessment.

Need greater emphasis on extended language, support for early literacy skills; also on high-quality free play and dramatic play, more rating of frequency/quality of reading aloud. No attention to issues of ELL students.

Overemphasis on health and safety, which seem to me to be the province of licensing; also, could be module on special needs children and handling them in the regular classroom, but they should be their own module.

Quality is such a broad term. Certainly the health and safety aspects are important and necessary, but we also need to pay more attention to teacher planning and assessment practices.

Some indicators reflect ideas that reflect a specific educational philosophy (e.g., about art activities).

[Missing indicators include those related to] staff characteristics such as training and education; staff job characteristics such as wages and benefits, working conditions.

The instrument talks as if activities are separate from a curriculum.

The standards are very much based on the first NAEYC guidelines for developmentally appropriate practice; they overemphasize time for individual privacy, gross motor/outdoor and dramatic play while under-emphasizing the time for informal math activities, hands on/teacher guided inquiry, children's explorations of books, the development of early literacy skills.

There are not clear delineations between structural constructs and interactional constructs that form a higher level of quality. The use of the language/reasoning in addition to activities and interaction causes some confusion.

There is a clear subset of items that deal with health, cleanliness, and safety. I think these should be pulled out as a separate index. You cannot say these are not important but they represent a different aspect. Similarly, I do not think computers or TVs should be included -- hard to rate because their inclusion at all is one I do not agree with.

There should be more items on staff monitoring and assessing children

These [indicators that are absent] include processes of child assessment, implementation of curriculum, leadership and administration, and more detailed evaluation of interaction between teachers and children.

Why is there so much talk about TV watching?

REFERENCES

Abbott-Shim, M., & Sibley, A. (1987). *Assessment profile for early childhood programs: Preschool, infant and school age*. Atlanta, GA: Quality Assist.

Arnett, J. (1989). Caregivers in day-care centers: Does training matter? *Journal of Applied Developmental Psychology, 10*, 541-442.

Bagnato, S. J. (2002, October). *Quality Early Learning--Key to School Success: A First-Phase 3-Year Program Evaluation Research Report for Pittsburgh's Early Childhood Initiative (ECI)*. Retrieved May 2, 2007, from the UCLID Center at the University of Pittsburgh's Web site: http://www.uclid.org:8080/uclid/ pdfs/ecp_final_report.pdf

Beller, E. K., Stahnke, M., Butz, P., Stahl, W., & Wessels, H. (1996). Two measures of the quality of group care for infants and toddlers. *European Journal of Psychology of Education, 6* (2), 151-167.

Belsky, J., Vandell, D. L., Burchinal, M., Clarke-Stewart, K. A., McCartney, K., & Owen, M. T. (2007, March/April). Are there long-term effects of child care? *Child Development, 78* (2), 681-701.

Bradley, R. H., Caldwell, B. M., Fitzgerald, J. A., Morgan, A. G., & Rock, S. L. (1996). Experience in day care and social competence among maltreated children. *Child Abuse and Neglect, 10,* 181-189.

Bredekamp, S. (1986). The reliability and validity of the Early Childhood Classroom Observation Scale for accrediting early childhood programs. *Early Childhood Research Quarterly, 1*, 103-118.

Bryant, D., Clifford, R. M., Saluja, G., Pianta, R., Early, D., Barbarin, O., et al. (2002). *Diversity and directions in state pre-kindergarten programs*. Chapel Hill: The University of North Carolina, FPG Child Development Institute, NCEDL.

Bryant, D., Maxwell, K., Taylor, K., Poe, M., Peisner-Feinberg, E., & Bernier, K. (2003). Smart Start and preschool child care quality in North Carolina: Change over time and relation to children's readiness. Chapel Hill, NC: FPG Child Development Institute.

Burchinal, M. R., Roberts, J. E., Riggins, R., Zeisel, S. A., Neebe, E., & Bryant, D. (2000, March/April). Relating quality of center-based child care to early cognitive and language development longitudinally. *Child Development, 71* (2), 339-357.

Buysse, V., Wesley, P. W. Bryant, D., & Gardner, D. (1999). Quality of early childhood programs in inclusive and noninclusive settings. *The Council for Exceptional Children, 65* (3), 301-314.

The Carolina Abecedarian Project (n.d.). Retrieved May 3, 2007, from http://www.fpg. unc.edu/~abc/

Cassidy, D. J., Hestenes, L. L., Hedge, 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* (3), 345-360.

Clifford, R. M. (2004, September). *Structure and stability of the Early Childhood Environment Rating Scale*. Paper presented at the meeting of the Center for Early Childhood Development and Education International Conference on Questions of Quality, Dublin, Ireland. Retrieved May 2, 2007, from http://www.cecde.ie/ english/ pdf/Questions %20of%20Quality/Clifford.pdf

Connor, C. M., Son, S., Hindman, A. H., & Morrison, F. J. (2005, October). Teacher qualifications, classroom practices, family characteristics, and preschool experience: Complex effects on first graders' vocabulary and early reading outcomes. *Journal of School Psychology, 43* (4), 343-375.

Crano, W. D., & Brewer, M. B. (2002). *Principles and Methods of Social Research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Cryer, D. (1999, May). Defining and assessing early childhood program quality. *The ANNALS of the American Academy of Political and Social Science, 563*, 39-55.

Cryer, D., Harms, T., & Riley, C. (2003). *All About the ECERS-R*. Lewisville, NC: Pact House Publishing.

Early, D. M., Maxwell, K. L., Burchinal, M., Alva, S., Bender, R. H., Bryant, D., et al. (2007, April). Teachers' education, classroom quality, and young children's academic skills: Results from seven studies of preschool programs. *Child Development, 78* (2), 558-580.

Environment Rating Scales (n.d.). Retrieved April 20, 2007, from the University of North Carolina Frank Porter Graham Child Development Institute Web site: http://www.fpg.unc.edu/~ecers/

Farran, D. C., Lipsey, M. W., Hurley, S., & Bilbrey, C. (2006, June). *The predictive utility of the ECERS_R in rural public school prekindergarten programs.* Poster session presented at the Head Start's Eighth National Research Conference, Washington, DC.

Gore, A. (7 January 1998). *Child Care Announcement.* [Transcript] Retrieved February 26, 2008, from http://clinton4.nara.gov/WH/EOP/OVP/speeches/chicare.html

Harms, T., & Clifford, R. M. (1980). *The Early Childhood Environment Rating Scale*. New York: Teachers College Press.

Harms, T., Clifford, R. M., & Cryer, D. (1998). *Early Childhood Environment Rating Scale* (Rev. ed.). Williston, VT: Teachers College Press.

Harms, T., Cryer, D., & Clifford, R. M. (1990). *Infant/Toddler Environment Rating Scale*. New York, NY: Teachers College Press.

Horne, S. (1984). Criterion-referenced testing: Pedagogical implications. *British Educational Research Journal, 10* (2), 155-173.

Howes, C., & Stewart, P. (1987). Child's play with adults, toys, and peers: an examination of family and child care influences. *Developmental Psychology, 23,* 423-430.

Hyson, M. C., Hirsh-Pasek, K. & Rescorla, L. (1990). The classroom practices inventory: An observation instrument based on National Association for the Education of Young Children's (NAEYC) guidelines for developmentally appropriate practices for 4- and 5-year-old children. *Early Childhood Research Quarterly, 5,* 475-494.

Kamerman, S. B., & Gatenio, S. (2003). Overview of the current policy context. In D. Cryer & R. M. Clifford (Eds.), *Early Childhood Education & Care in the USA* (pp. 1-30). Baltimore: Brookes.

Kirp, D. L. (2007). *The Sandbox Investment: The preschool movement and kids-first politics.* Cambridge, MA: Harvard University Press.

Kontos, S., Burchinal, M., Howes, C., Wisseh, S., & Galinsky, E. (2002). An eco-behavioral approach to examining the contextual effects of early childhood classrooms. *Early Childhood Research Quarterly, 17* (2), 239-258.

LaParo, K. M., & Pianta, R. C. (2003). *CLASS: Classroom Assessment Scoring System*. Charlottesville: University of Virginia.

LaParo, K. M., Pianta, R. C., & Stuhlman, M. (2004, May). The classroom assessment scoring system: findings from the prekindergarten year. *The Elementary School Journal, 104* (5), 409-426.

Lee, J. & Walsh, D. J. (2004). Quality in early childhood programs: Reflections from program evaluation practices. *American Journal of Evaluation, 25* (3), 351-373.

Melhuish, E. C. (2001). The quest for quality in early day care and preschool experience continues. *International Journal of Behavioral Development, 25* (1), 1-6.

National Association for the Education of Young Children. (n.d.). *NAEYC Academy for Early Childhood Program Accreditation*. Retrieved March 5, 2008, from http://www.naeyc.org/academy/

National Center for Early Development and Learning (1997). *Classroom Observation System—Kindergarten*. Charlottesville, VA: University of Virginia.

The National Institute for Early Education Research (2006). *The State of Preschool 2006: State Preschool Yearbook.* Retrieved April 28, 2007, from http://nieer.org/yearbook/pdf/yearbook.pdf.

NICHD Early Child Care Research Network (1996). Characteristics of infant child care: Factors contributing to positive caregiving. *Early Childhood Research Quarterly, 11,* 269-306.

Peisner-Feinberg, E. S., Burchinal, M. R., Clifford, R. M., Culkin, M. L., Howes, C., Kagan, S. L., et al. (2001, October). The relation of preschool child-care quality to children's cognitive and social developmental trajectories through second grade. *Child Development, 72* (5), 1534-1553.

Perlman, M., Zellman, G. L., & Le, V. (2004). Examining the psychometric properties of the Early Childhood Environment Rating Scale-Revised (ECERS-R). *Early Childhood Research Quarterly, 19*, 398-412.

Peterson, C., & Peterson, R. (1986). Parent-child interaction and day care: Does quality of day care matter? *Journal of Applied Developmental Psychology, 7,* 1-15.

Phillips, D., Mekos, D., Scarr, S., McCartney, K., & Abbott-Shim, M. (2000). Within and beyond the classroom door: Assessing quality in child care centers. *Early Childhood Research Quarterly, 15* (4), 475-496.

Phillipsen, L. C., Burchinal, M. R., Howes, C., & Cryer, D. (1997). The prediction of process quality from structural features of child care. *Early Childhood Research Quarterly, 12* (3), 281-303.

Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., et al. (2005). Features of pre-kindergarten programs, classrooms, and teachers: Do they predict observed classroom quality and child-teacher interactions? *Applied Developmental Science, 9* (3), 144-159.

Pianta, R. C., LaParo, K. M., & Hamre, B. K. (2006). *Classroom Assessment Scoring System Manual, Preschool (Pre-K) Version*. Charlottesville, VA: Center for Advanced Study of Teaching and Learning.

Pianta, R. C., La Paro, K. M., Payne, C., Cox, M. J., & Bradley, R. (2002, January). The relation of kindergarten classroom environment to teacher, family, and school characteristics and child outcomes. *The Elementary School Journal, 102* (3), 225-238.

Prescott, E., Kritchevsky, S., & Jones, K. (1972). *The day care environment inventory*. Washington DC: Department of Health, Education and Welfare.

Reeves, C. A., & Bednar, D. A. (1994, July). Defining quality: Alternatives and implications. *Academy of Management Review, 19* (3), 419-445.

Rimm-Kaufman, S. E., La Paro, K. M., Downer, J. T., & Pianta, R. C. (2005, March). The contribution of classroom setting and quality of instruction to children's behavior in kindergarten classrooms. *The Elementary School Journal, 105* (4), 377-395.

Ritchie, S., Howes, C., Kraft-Sayre, M., & Weiser, B. (2002). *Snapshot*. Los Angeles: University of California, Los Angeles.

Sado, S., & Bayer, A. (2001, June). *Executive summary: The changing American family*. Retrieved from the Population Resource Center Webs site http://www.prcdc.org/ summaries/family/family.html

Safe, Smart, and Happy Kids: Information for Tennessee's Licensed Child Care Providers About the Child Care Evaluation and Report Card and Star-Quality Programs (2005). Retreived May 3, 2007, from http://tnstarquality.org/

Sakai, L. M., Whitebook, M.,Wishard, A., & Howes, C. (2003). Evaluating the Early Childhood Environment Rating Scale (ECERS): Assessing differences between the first and revised edition. *Early Childhood Research Quarterly, 18,* 427–445.

Scarr, S., Eisenberg, M., & Deater-Deckard, K. (1994, June). Measurement of quality in child care centers. *Early Childhood Research Quarterly, 9* (2), 131-151.

Schaefer, S. (2003, February). *A child advocate's guide to federal early care and education policy.* Retrieved April 20, 2007, from Voices for America's Children Web site: http://www.voicesforamericaschildren.org/Content/ContentGroups/ Publications1/Voices_for_Americas_Children/ECE1/20031/ECEFederalGuide. pdf

Smith, K. E., & Bachu, A. (1999, January). *Women's labor force attachment patterns and maternity leave: A review of the literature* (Population Division Working Paper No. 32). Retrieved April 28, 2007, from U. S. Census Bureau Web site http://www.census.gov/population/www/documentation/twps0032/twps0032.html

Smith, M.W., & Dickinson, D. K., (with Sangeorge, A., & Anastasopoulos, L.) (2002). *Early Language & Literacy Classroom Observation. (*Research ed.). Baltimore, MD: Paul Brookes.

Step Up to Quality (n.d.). Retrieved May 3, 2007, from the Ohio Job and Family Services Web site: http://jfs.ohio.gov/cdc/stepUpQuality.stm

Stipek, D. (1996). *Early Childhood Classroom Observation Measure*. Los Angeles: University of California at Los Angeles, Graduate School of Education.

Stipek, D. & Byler, P. (2004). The early childhood classroom observation measure. *Early Childhood Research Quarterly, 19* (3), 375-397.

Stipek, D., Daniels, D. Galuzzo, D., & Milburn, S. (1992). Characterizing early childhood education programs for poor and middle-class children. *Early Childhood Research Quarterly, 7,* 1-19.

Stipek, D., Milburn, S., Clements, D., & Daniels, D. (1992). Parents' beliefs about appropriate education for young children. *Journal of Applied Developmental Psychology,* 13, 293-210.

Sylva, K., Siraj-Blatchford, I., & Taggart, B. (2003). *Assessing quality in the early years: Early Childhood Environment Rating Scale-Extension (ECERS-E): Four curricular subscales.* Stoke-on-Trent: Trentham Books.

Sylva, K., Siraj-Blatchford, I., Taggart, B., Sammons, P., Melhuish, E., Elliot, K., et al. (2006). Capturing quality in early childhood through environmental rating scales. *Early Childhood Research Quarterly, 21* (1), 76-92.

Tietze, W., Cryer, D., Bairrao, J., Palacios, J., & Wetzel, G. (1996). Comparisons of observed process quality in early child care and education programs in five countries. Early Childhood Research Quarterly, 11, 447-475.

U. S. Census Bureau (n.d.). *Current Population Survey Reports: Historical Tables- Table A-2. Percentage of the Population 3 Years Old and Over Enrolled in School, by Age, Sex, Race, and Hispanic Origin: October 1947 to 2005*. Retrieved April 28, 2007 from U. S. Census Web site http://www.census.gov/ population/ www/socdemo/school.html

U. S. Department of Education, National Center for Education Statistics (2006). *The Condition of Education 2006* (NCES 2006-071). Retrieved February 26, 2008, from http://nces.ed.gov/pubs2006/2006071.pdf

Van Horn, M. L., Karlin, E. O., Ramey, S. L., Aldridge, J., & Snyder, S. W. (2005, March). Effects of developmentally appropriate practices on children's development: A review of research and discussion of methodological and analytic issues. *The Elementary School Journal, 105* (4), 325-351.

Whitebook, M., Howes, C., & Phillips, D. (1990). *Who cares? Child care teachers and the quality of care in America: Executive summary, National Child Care Staffing Study*. Retrieved April 20, 2006, from http://www.ccw.org/publications_ archives.html

Wishard, A. G., Shivers, E. M., Howes, C., & Ritchie, S. (2003). Child care program and teacher practices: associations with quality and children's experiences. *Early Childhood Research Quarterly, 18* (1), 65-103.

World Health Organisation (1990). *WHO child care facility schedule with user's manual*. Geneva: WHO Division of Mental Health.

Wolfe, J. (2002). *Learning from the Past: Historical Voices in Early Childhood Education* (Rev. 2nd ed.). Mayerthorpe, Alberta: Piney Branch Press.